# Stochastic neighbor embedding (SNE) for dimension reduction and visualization using arbitrary divergences

Bunte, Kerstin; Haase, Sven; Biehl, Michael; Villmann, Thomas

*Published in:*
Neurocomputing

*DOI:*
10.1016/j.neucom.2012.02.034



*Publication date:*
2012

Neurocomputing 90 (2012) 23–45

Contents lists available at SciVerse ScienceDirect

Neurocomputing

journal homepage: www.elsevier.com/locate/neucom

# Stochastic neighbor embedding (SNE) for dimension reduction and visualization using arbitrary divergences

Kerstin Bunte [a,b,*], Sven Haase [c], Michael Biehl [a], Thomas Villmann [c]

[a] Johann Bernoulli Institute for Mathematics and Computer Science, University of Groningen, P.O. Box 407, 9700AK Groningen, The Netherlands
[b] University of Bielefeld, CITEC Center of Excellence, D-33501 Bielefeld, Germany
[c] Department of Mathematics, University of Applied Sciences Mittweida, Germany

ARTICLE INFO

Available online 20 March 2012

*Keywords:*
Dimension reduction
Visualization
Divergence optimization
Nonlinear embedding
Stochastic neighbor embedding

ABSTRACT

We present a systematic approach to the mathematical treatment of the t-distributed stochastic neighbor embedding (t-SNE) and the stochastic neighbor embedding (SNE) method. This allows an easy adaptation of the methods or exchange of their respective modules. In particular, the divergence which measures the difference between probability distributions in the original and the embedding space can be treated independently from other components like, e.g. the similarity of data points or the data distribution. We focus on the extension for different divergences and propose a general framework based on the consideration of Fréchet-derivatives. This way the general approach can be adapted to the user specific needs.



## 1. Introduction

Various dimension reduction techniques have been introduced based on the aim of preserving specific properties of the original data. The spectrum ranges from linear projections of original data, such as principal component analysis (PCA) or classical multi-dimensional scaling (MDS) [1] to a variety of locally linear and nonlinear approaches, such as isomap [2,3], locally linear embedding (LLE) [4], local linear coordination (LLC) [5], or charting [6,7].

Other methods aim at the preservation of the classification accuracy in lower dimensions and incorporate the available label information for the embedding, e.g. linear discriminant analysis (LDA) [8] and generalizations thereof [9], extensions of the self-organizing map (SOM) [10], incorporating class labels [11], and limited rank matrix learning vector quantization (LiRaM LVQ) [12,13]. For a comprehensive review on nonlinear dimensionality reduction methods, we refer to [14].

Recently, the stochastic neighbor embedding (SNE) [15] and extensions thereof have become popular for visualization. SNE approximates the probability distribution in the high-dimensional space, defined by neighboring points, with their probability distribution in a lower-dimensional space. In [16] the authors proposed a technique called t-SNE, which is a variation of SNE considering a particular statistical model assumption for data distributions. The similarity of the distributions is quantified in terms of the Kullback–Leibler divergence. In [17] it is argued that the preservation of shift-invariant similarities as employed by SNE and its variants is superior in comparison to distance preservation as performed by many traditional dimension reduction techniques.

Functional metrics like Sobolev distances, kernel-based dissimilarity measures and divergences have attracted attention recently for the processing of data showing a functional structure. These dissimilarity measures were for example investigated as alternatives to the most common choice, the Euclidean distance [18–22]. The application of divergences for Vector Quantization and Learning Vector Quantization schemes have been investigated in [23,24].

This work bases on [25], where the self-organized neighbor embedding (SONE), which can be seen as a hybrid between the self-organizing map (SOM) and SNE, has been extended to the use of arbitrary divergences. In this contribution, we formulate a mathematical framework based on Fréchet derivatives which allows to generalize the concept of SNE and t-SNE to arbitrary divergences. This leads to a new dimension reduction and visualization scheme, which can be adapted to the user specific requirements in an actual problem. We summarize the general classes of divergences following the scheme introduced by [26] and extended in [23]. The mathematical framework for functional derivatives of continuous divergences is given by the functional-analytic generalization of common derivatives, known as Fréchet derivatives [27,28]. It is the generalization of partial derivatives for the discrete variants of the divergences.

* Corresponding author at: CITEC Center of Excellence - Cognitive Interaction Technology, Bielefeld University, Universitätsstraße 21–23, D-33615 Bielefeld, Germany. Tel.: +49 521106 12130.
*E-mail address:* kerstin.bunte@googlemail.com (K. Bunte).
*URL:* http://www.cit-ec.de/de/tcs/kerstin (K. Bunte).

0925-2312/$ - see front matter © 2012 Elsevier B.V. All rights reserved.
http://dx.doi.org/10.1016/j.neucom.2012.02.034

We introduce a general mathematical framework for the extension of SNE and t-SNE for arbitrary divergences. The different classes of divergences are characterized and for various examples the Fréchet derivatives are identified. We demonstrate the proposed framework for the example case of the Gamma divergence. The behavior of different divergences stemming from the identified divergence families are shown on several examples in the image analysis domain.

## 2. Review of SNE and t-SNE

Generally, dimensionality reduction methods convert a high dimensional data set $\{x_i\}_{i=1}^n \in \mathbb{R}^N$ into low dimensional data $\{\xi_i\}_{i=1}^n \in \mathbb{R}^M$. A probabilistic approach to visualize the structure of complex data sets, preserving neighbor similarities is stochastic neighbor embedding (SNE), proposed by Hinton and Roweis [15]. SNE converts high-dimensional Euclidean distances between data points into probabilities that represent similarities. The conditional probabilities $p_{j|i}$ that a data point $x_i$ would pick $x_j$ as its neighbor is given by

$$p_{j|i} = \frac{\exp(-\|x_i - x_j\|^2/2\sigma_i^2)}{\sum_{j\neq i}\exp(-\|x_i - x_j\|^2/2\sigma_i^2)}, \tag{1}$$

with $p_{i|i} = 0$. The variance $\sigma_i$ of the Gaussians centered around $x_i$ is determined by a binary search procedure [16]. The density of the data is likely to vary. In dense regions a smaller value of $\sigma$ is more appropriate than in sparse regions. Let $P_i$ be the conditional probability distribution over all other data points given point $x_i$. This distribution has an entropy which increases as $\sigma_i$ increases. SNE performs a binary search for the value of $\sigma_i$ which produces a $P_i$ with a fixed perplexity specified by the user. The perplexity is defined as

$$\mathrm{perpl}(P_i) = 2^{H(P_i)}, \tag{2}$$

where $H(P_i)$ is the Shannon entropy of $P_i$ measured in bits: $H(P_i) = -\sum_j p_{j|i}\log_2 p_{j|i}$. It can be interpreted as a smooth measure of the effective number of neighbors and typical values ranges between 5 and 50 dependent on the data set size.

The low-dimensional counterparts $\xi_i$ and $\xi_j$ of the high-dimensional data points $x_i$ and $x_j$ are modeled by similar probabilities

$$q_{j|i} = \frac{\exp(-\|\xi_i - \xi_j\|^2)}{\sum_{j\neq i}\exp(-\|\xi_i - \xi_j\|^2)}, \tag{3}$$

with again $q_{i|i} = 0$. SNE tries to find a low-dimensional data representation which minimizes the mismatch between the conditional probabilities $p_{j|i}$ and $q_{j|i}$. As a measure of mismatch the Kullback–Leibler divergence $\mathrm{D_{KL}}$ is used such that the cost function SNE is given by

$$C = \sum_i \mathrm{D_{KL}}(P_i\|Q_i) = \sum_i\sum_j p_{j|i}\log\frac{p_{j|i}}{q_{j|i}}, \tag{4}$$

where $Q_i$ is defined similar to $P_i$ as the conditional probability distribution over all other points given $\xi_i$. The cost function is not symmetric and focuses on retaining the local structure of the data in the mapping. Large costs appear for mapping nearby data points widely separated in the embedding, but there is only small cost for mapping widely separated data points close together. The minimization of the cost function equation (4) is performed using a gradient descent approach. For details we refer to [15].

The so-called "crowding problem" may be observed in SNE and other local techniques, like for example Sammon mapping [16]. The (even very small) attractive forces might crush together moderately dissimilar points in the center of the map. Therefore, in [16] van der Maaten and Hinton presented a technique called t-SNE, which is a variation of SNE considering another statistical model assumption for the data distribution to avoid that problem. Instead of using the conditional probabilities $p_{j|i}$ and $q_{j|i}$ the joint probability distributions $P$ and $Q$ are used to optimize a symmetric version of SNE with the cost function

$$C = \mathrm{D_{KL}}(P\|Q) = \sum_i\sum_j p_{ij}\log\frac{p_{ij}}{q_{ij}} \tag{5}$$

with $p_{ii} = q_{ii} = 0$. Here, the pairwise similarities in the high-dimensional space are defined by the conditional probabilities

$$p_{ij} = \frac{p_{j|i} + p_{i|j}}{2n} \tag{6}$$

and the low-dimensional similarities are given by

$$q_{ij} = \frac{(1+\|\xi_i - \xi_j\|^2)^{-1}}{\sum_{k\neq l}(1+\|\xi_k - \xi_l\|^2)^{-1}}. \tag{7}$$

The application of the heavy-tailed Student $t$-distribution with one degree of freedom allows to model moderate distances in the high-dimensional space by much larger distances in the embedding. Therefore, the unwanted attractive forces between map points that represent moderately dissimilar data points is eliminated. See [16] for further details.

## 3. A generalized framework

In this paper we provide the mathematical framework for the generalization of t-SNE and SNE, with respect to the use of arbitrary divergences in the cost-function for the gradient descent. We generalize the definitions towards continuous measures in the high-dimensional space $\mathcal{X} = \{x,y\}$ and a low-dimensional space $\mathcal{E} = \{\xi,\zeta\} \in \mathbb{R}^M$. The pairwise similarities in the high-dimensional original data space are set to

$$p = p_{xy} = \frac{p_{y|x} + p_{x|y}}{2\cdot\int 1\, dy'}, \tag{8}$$

with conditional probabilities

$$p_{y|x} = \frac{\exp(-\|x-y\|^2/2\sigma_x^2)}{\int\exp(-\|x-y'\|^2/2\sigma_x^2)\, dy'}.$$

### 3.1. The generalized t-SNE gradient

Let $\mathrm{D}(p\|q)$ be a divergence for non-negative integrable measure functions $p = p(r)$ and $q = q(r)$ with a domain $V$ and $\xi,\zeta \in \mathcal{E}$ distributed according to $\Pi_\mathcal{E}$ [26]. Further, let $r(\xi,\zeta) : \mathcal{E}\times\mathcal{E}\to\mathbb{R}$ with the distribution $\Pi_r = \phi(r,\Pi_\mathcal{E})$. Let us use the squared Euclidean distance in the low dimensional space:

$$r = r(\xi,\zeta) = \|\xi - \zeta\|^2. \tag{9}$$

For t-SNE, $q$ is obtained by means of a Student $t$-distribution, such that

$$q(r(\xi',\zeta')) = \frac{(1+r(\xi',\zeta'))^{-1}}{\iint(1+r(\xi,\zeta))^{-1}\, d\xi\, d\zeta}, \tag{10}$$

which we will abbreviate below for reasons of clarity as

$$q(r') = \frac{(1+r')^{-1}}{\int\int(1+r)^{-1}\, d\xi\, d\zeta} = f(r')\cdot I^{-1}. \tag{11}$$

Now let us consider the derivative of D with respect to $\xi$:

$$\begin{aligned}\frac{\partial \mathrm{D}}{\partial\xi} &= \frac{\partial \mathrm{D}(p,q(r(\xi,\zeta)))}{\partial\xi} = \iint\frac{\delta \mathrm{D}}{\delta r'}\frac{\partial r'}{\partial\xi}\, d\xi'\, d\zeta' \\ &= \iint\frac{\delta \mathrm{D}}{\delta r(\xi',\zeta')}\frac{\partial r(\xi',\zeta')}{\partial\xi}\, d\xi'\, d\zeta' = 4\int\frac{\delta \mathrm{D}}{\delta r(\xi,\zeta)}(\xi-\zeta)\, d\zeta.\end{aligned} \tag{12}$$

We now have to consider $\delta D/\delta r(\xi,\zeta)$. Again, using the chain rule for functional derivatives we get

$$\frac{\delta D}{\delta r(\xi,\zeta)} = \iint \frac{\delta D}{\delta q(r(\xi',\zeta'))} \frac{\delta q(r(\xi',\zeta'))}{\delta r(\xi,\zeta)}\, d\xi'\, d\zeta' = \int \frac{\delta D}{\delta q(r')} \frac{\delta q(r')}{\delta r} \Pi_{r'}\, dr', \tag{13}$$

where

$$\frac{\delta q(r')}{\delta r} = \frac{\delta f(r')}{\delta r} \cdot I^{-1} - f(r') \cdot I^{-2} \frac{\delta I}{\delta r}$$

holds, with

$$\frac{\delta f(r')}{\delta r} = -\delta_{r,r'}(1+r)^{-2} \quad \text{and} \quad \frac{\delta I}{\delta r} = -(1+r)^{-2}.$$

So we obtain

$$\begin{aligned}\frac{\delta q(r')}{\delta r} &= f(r') \cdot I^{-2} \cdot \frac{1}{(1+r)^2} - \frac{\delta_{r,r'}(1+r)^{-2}}{I} \\ &= \frac{f(r')}{I}\frac{f(r)}{I}\frac{1}{(1+r)} - \frac{\delta_{r,r'}(1+r)^{-1} f(r)}{I} \\ &= q(r')q(r)\frac{1}{(1+r)} - \delta_{r,r'}(1+r)^{-1} q(r) \\ &= -(1+r)^{-1} q(r)(\delta_{r,r'} - q(r')).\end{aligned}$$

Substituting these results in Eq. (13), we get

$$\begin{aligned}\frac{\delta D}{\delta r} &= \int \frac{\delta D}{\delta q(r')} \frac{\delta q(r')}{\delta r} \Pi_{r'}\, dr' \\ &= -\frac{q(r)}{1+r} \int \frac{\delta D}{\delta q(r')} (\delta_{r,r'} - q(r')) \Pi_{r'}\, dr' \\ &= -\frac{q(r)}{1+r} \left( \frac{\delta D}{\delta q(r)} - \int \frac{\delta D}{\delta q(r')} q(r') \Pi_{r'}\, dr' \right).\end{aligned}$$

Finally collect all terms and get

$$\begin{aligned}\frac{\partial D}{\partial \xi} &= 4 \int \frac{\delta D}{\delta r} (\xi - \zeta)\, d\zeta \\ &= 4 \int \frac{-q(r)}{1+r} \left[ \frac{\delta D}{\delta q(r)} - \int \frac{\delta D}{\delta q(r')} q(r') \Pi_{r'}\, dr' \right] \cdot (\xi - \zeta)\, d\zeta.\end{aligned} \tag{14}$$

We now have the obvious advantage that we can derive $\partial D/\partial \xi$ for several divergences $D(p\|q)$ directly from Eq. (14), if the Fréchet derivative $\delta D/\delta q(r)$ of D with respect to $q(r)$ is known. The concept of Fréchet derivatives and explicit formulas for different divergences are given in Section 6.

### 3.2. The generalized SNE gradient

In symmetric SNE, the pairwise similarities in the low dimensional map are given by [16]

$$q_{SNE}{}' = q_{SNE}(r(\xi',\zeta')) = \frac{\exp(-r(\xi',\zeta'))}{\iint \exp(-r(\xi,\zeta))\, d\xi\, d\zeta},$$

which we will abbreviate below for reasons of clarity as

$$q_{SNE}(r') = \frac{\exp(-r')}{\iint \exp(-r)\, d\xi\, d\zeta} = g(r') \cdot J^{-1}, \tag{15}$$

with $g(r') = \exp(-r')$ and $J$ representing the integral in the denominator. Consequently, if we consider $\partial D/\partial \xi$, we can use the results from above for t-SNE. The only term that differs is the derivative of $q_{SNE}(r')$ with respect to $r$. Therefore we get

$$\frac{\delta q_{SNE}(r')}{\delta r} = \frac{\delta g(r')}{\delta r} \cdot J^{-1} - g(r') \cdot J^{-2} \frac{\delta J}{\delta r},$$

with

$$\frac{\delta g(r')}{\delta r} = -\delta_{r,r'} \exp(-r) \quad \text{and} \quad \frac{\delta J}{\delta r} = -\exp(-r),$$

which leads to

$$\begin{aligned}\frac{\delta q_{SNE}(r')}{\delta r} &= \frac{-\delta_{r,r'} \exp(-r)}{J} + g(r') J^{-2} \exp(-r) \\ &= \frac{-\delta_{r,r'} g(r)}{J} + \frac{g(r')}{J}\frac{g(r)}{J} = -\delta_{r,r'} q_{SNE}(r) + q_{SNE}(r') q_{SNE}(r) \\ &= -q_{SNE}(r)(\delta_{r,r'} - q_{SNE}(r')).\end{aligned}$$

Substituting these results in Eq. (13), we get

$$\begin{aligned}\frac{\delta D}{\delta r} &= \int \frac{\delta D}{\delta q_{SNE}(r')} \frac{\delta q_{SNE}(r')}{\delta r} \Pi_{r'}\, dr' \\ &= -q_{SNE}(r) \cdot \int \frac{\delta D}{\delta q_{SNE}(r')} (\delta_{r,r'} - q_{SNE}(r')) \Pi_{r'}\, dr' \\ &= -q_{SNE}(r) \cdot \left[ \frac{\delta D}{\delta q_{SNE}(r)} - \int \frac{\delta D}{\delta q_{SNE}(r')} q_{SNE}(r') \Pi_{r'}\, dr' \right].\end{aligned}$$

Finally, substituting this result in Eq. (12), we obtain

$$\begin{aligned}\frac{\partial D}{\partial \xi} &= 4 \int \frac{\delta D}{\delta r} (\xi - \zeta)\, d\zeta \\ &= -4 \int q_{SNE}(r)(\xi - \zeta) \cdot \left[ \frac{\delta D}{\delta q_{SNE}(r)} - \int \frac{\delta D}{\delta q_{SNE}(r')} q_{SNE}(r') \Pi_{r'}\, dr' \right] d\zeta\end{aligned} \tag{16}$$

as the general formulation of the SNE cost function gradient, which uses the Fréchet-derivatives of the applied divergences as above for t-SNE.

## 4. Specifications of divergences

Divergences are derived from simple component-wise errors, e.g. the Euclidean and Minkowski metrics [26]. These frequently used metrics are intuitive and they are optimal estimators in case of Gaussian noise or error. However, if the observations are corrupted not only by Gaussian noise but also by outliers, estimators based on these metrics can be strongly biased. They also suffer from the curse of dimensionality, which means that observations become equidistant in terms of the Euclidean distance for high-dimensional data. In many applications like pattern matching, image analysis, statistical learning, etc. the noise is not necessarily Gaussian and information divergences are used. Employing generalized divergences might provide a compromise between the efficiency and robustness and/or compromise between a mean squared error and bias.

Divergences are functionals $D(p\|q)$ designed as dissimilarity measures between two non-negative integrable functions $p$ and $q$ [26]. In practice, usually $p$ corresponds to the observed data and $q$ denotes the estimated or expected data. We assume $p(r)$ and $q(r)$ are positive measures defined on $r$ in the domain $V$. The weight of the functional $p$ is defined as

$$W(p) = \int_V p(r)\, dr. \tag{17}$$

Positive measures with the additional constraint $W(p) = 1$ can be interpreted as probability density functions. Generally speaking, divergences measure a quasi-distance or directed difference, while we are mostly interested in separable measures, which satisfy the condition

$$D(p\|q) \begin{cases} > 0 & \text{for } p \neq q, \\ = 0 & \text{iff } p \equiv q. \end{cases} \tag{18}$$

In contrast to a metric, divergences may be non-symmetric $D(p\|q) \neq D(q\|p)$, and do not necessarily satisfy the triangular inequality $D(p\|q) \leq D(p\|z) + D(z\|q)$. Following [26] one can distinguish at least three main families of divergences with the same consistent properties: Bregman-divergences, Csiszár's $f$-divergences and $\gamma$-divergences. Note that all these families contain the Kullback–Leibler (KL)

divergence as a special case, so the KL-divergence can be seen as the non-empty intersection between the sets of divergences.

In general we assume $p$ and $q$ to be positive measures. In case they are normalized we refer to them as probability densities. We review some basic properties of divergences in the following sections. For detailed information we refer to [26,29].

An overview of the family of divergences, examples and their relationship to each other can be found in Fig. 1. Some important properties are summarized in Tables 1 and 2. We review the families of divergences and some examples in the following sections.

### 4.1. Bregman divergences

A Bregman divergence is defined as a pseudo-distance between two positive measures $p$ and $q$: $D_B(p\|q): \mathcal{L}\times\mathcal{L}\to\mathbb{R}^+$. Let $\phi$ be a strictly convex real-valued function with the domain of the Lebesgue-integrable functions $\mathcal{L}$ and twice continuously Fréchet-differentiable [28]. Then the Bregman divergence can be defined by

$$D_B^{\phi}(p\|q) = \int \phi(p) - \phi(q) - \frac{\delta\phi(q)}{\delta q}[p-q]\,dr, \tag{19}$$

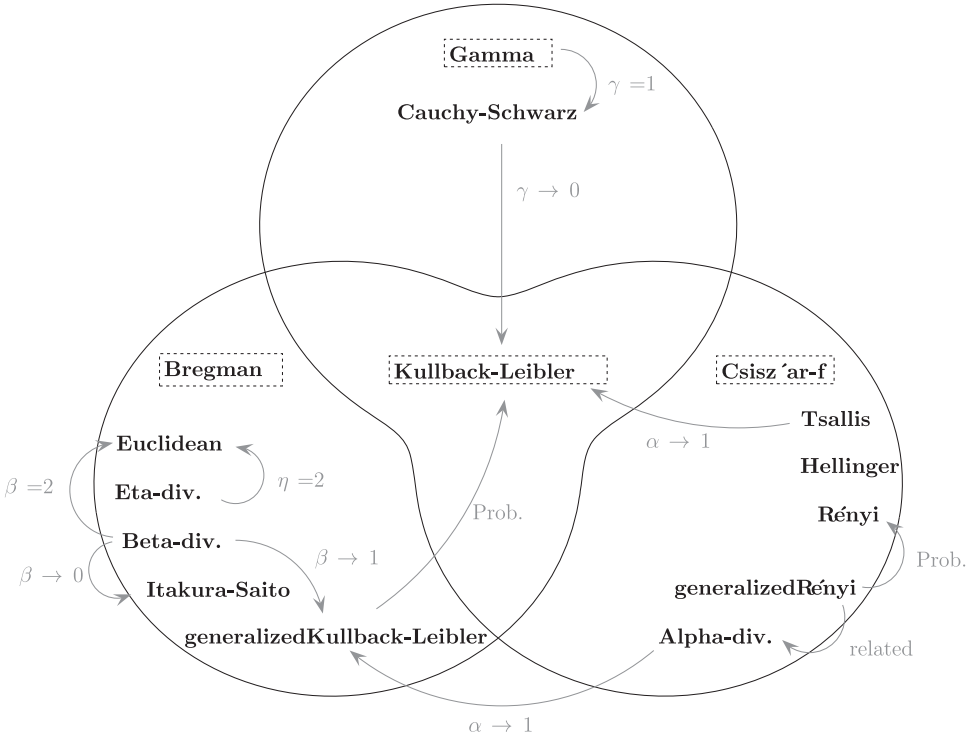


**Fig. 1.** Overview over the families of divergences and their relationship to each other. The shortcut *Prob.* denotes the special case of probability densities. For the sake of clarity we show the most important relations only and do not claim completeness.

**Table 1**
Table of divergences and their properties. The example divergences inherit the properties of the divergence family (gray box) and sometimes they show additional properties, stated individually. The shortcut (*pd*) denotes that the divergence is defined only for probability densities.

| Divergence (generating function) | (most) Important properties | | | | | |
|---|---|---|---|---|---|---|
| Bregman | Entropy | Convexity in $p$ | Linearity | Invariance affine transf. | Three-point property | Pythagoras Theorem |
| $D_B^{\phi}(p\|q) = [\int \phi(p) - \phi(q) - \frac{\delta\phi(q)}{\delta q}[p-q]\,dr]$ | $H_{\Phi}(p) = -\int \Phi(p)\,dr$ | | | | | |
| Gen. Kullback–Leibler | Shannon Entropy | | | | | |
| $[\Phi(u) = \int (u\log u - u)\,dr]$ | $H_S(p) = -\int p\ln(p)$ | | | | | |
| Itakura–Saito | Burg Entropy | | | | | |
| $[\Phi(u) = -\int \log u\,dr]$ | $H_B(p) = \int(p)$ | | | | | |
| Eta-divergence | | | | | | |
| $[\Phi(u) = \int u^{\eta}\,dr,\ \eta > 1]$ | | | | | | |
| Beta-divergence | Related to | | Scaling | Rescaled Eta-div. | | |
| $\left[\Phi(u) = \frac{u^{\beta} - \beta\cdot u + \beta - 1}{\beta(\beta-1)}\right]$ | Tsallis Entropy | | $D_{\beta}(cp\|cq) = c^{\beta} D_{\beta}(p\|q)$ | | | |
| Euclidean | | | | | Symmetric | |
| $[\Phi(u) = u^2]$ | | | | | | |
| Gamma divergence | Related to Rényi Entropy | | Scale invariant | | | Pythagoras Theorem |
| | | | $D_{\gamma}(cp\|cq) = D_{\gamma}(p\|q)$ | | | |
| Cauchy–Schwarz | | | | | Symmetric | Cauchy–Schwarz inequality |
| $\gamma = 1$ | | | | | | |

**Table 2**
Table of divergences and their properties (continued).

| Divergence | (most) Important properties | | | | | |
|---|---|---|---|---|---|---|
| Gen. Csiszár-f<br>$D_f^G(p\|q)=c_f\int(p-q)\,dr+\int qf\left(\frac{p}{q}\right)dr$ | Gen. Entropy<br>$H_f(p)=-\int f(p)\,dr$ | Convexity to both $p$, $q$ | Scaling<br>$cD_f=D_{cf}, c>0$ | Invariance<br>bijective transf. | Symmetry<br>$f_{\mathrm{sym}}(u)=f(u)+f^*(u)$ | Upper bound |
| Csiszár f divergence ($pd$)<br>$D_f(p\|q)=\int q\cdot f\left(\frac{p}{q}\right)dr$ | Generalized Entropy<br>$H_f(p)=-\int f(p)\,dr$ | Convexity | Scaling<br>$cD_f=D_{cf}, c>0$ | Invariance<br>bijective transf. | Symmetry<br>$f_{\mathrm{sym}}(u)=f(u)+f^*(u)$ | Bounded |
| Alpha divergence<br>$f(\frac{p}{q})=\frac{p}{q}\frac{\left(\frac{p}{q}\right)^{\alpha-1}-1}{\alpha^2-\alpha}+\frac{1-\frac{p}{q}}{\alpha}$ | Related to Tsallis Entropy | Convexity to both $p$, $q$ | Scaling<br>$D_\alpha(cp\|cq)=cD_\alpha(p\|q)$ | Duality<br>$D_\alpha=D_{1-\alpha}$ | Continuity | |
| Hellinger ($pd$)<br>$\left[f\left(\frac{p}{q}\right)=2\left(1-\sqrt{\frac{p}{q}}\right)\right]$ | | | | | | |
| Tsallis ($pd$) | Tsallis Entropy | | | Rescaled Alpha div. | | |

where $\delta\phi(q)/\delta q$ is the Fréchet derivative of $\phi$ with respect to $q$ [23]. Well known fundamental properties of the Bregman divergences are [26]:

*Convexity*: A Bregman divergence is always convex in its first argument but not necessary in its second.

*Non-negativity*:

$$D_B^\phi(p\|q)\geq 0 \quad \text{and} \quad D_B^\phi(p\|q)=0 \text{ iff } p\equiv q. \tag{20}$$

*Linearity*: Bregman divergences are linear according to the generating function $\Phi$. Any positive linear combination of Bregman divergences is also a Bregman divergence:

$$D_B^{c_1\phi_1+c_2\phi_2}(\cdot)=c_1D_B^{\phi_1}(\cdot)+c_2D_B^{\phi_2}(\cdot), \quad c_1,c_2>0.$$

*Invariance*: A Bregman divergence is invariant under affine transformations. Thus, $D_B^\Gamma(p\|q)=D_B^\phi(p\|q)$ is valid for any affine transformation

$$\Gamma(q)=\phi(q)+\Psi_g[q]+c, \tag{21}$$

with linear operator

$$\Psi_g[q]=\frac{\delta\Gamma(g)}{\delta g}\cdot q-\frac{\delta\phi(g)}{\delta g}\cdot q \tag{22}$$

for positive measures $g$ and $q$ and scalar $c$.

*Three-point property*: For any triple $p,q,g$ of positive measures the property holds:

$$D_B^\phi(p\|g)=D_B^\phi(p\|q)+D_B^\phi(q\|g)+(p-q)\left(\frac{\delta\phi(q)}{\delta q}-\frac{\delta\phi(g)}{\delta g}\right). \tag{23}$$

*Generalized Pythagorean theorem*: Let $P_\Omega(q)=\arg\min_{\omega\in\Omega}D_B^\phi(\omega\|q)$ be the Bregman projection onto the convex set $\Omega$ and $p\in\Omega$. The inequality

$$D_B^\phi(p\|q)\geq D_B^\phi(p\|P_\Omega(q))+D_B^\phi(P_\Omega(q)\|q) \tag{24}$$

is known as generalized Pythagorean theorem. If $\Omega$ is an affine set it holds with equality.

*Optimality*: In [30] an optimality property is stated. Given a set $S$ of positive measures $p$ with mean $\mu=E[S]$ and $\mu\in S$ the unique minimizer $E_{p\in S}[D(p\|q)]$ is minimum for $q=\mu$ if D is a Bregman divergence. This property favors the Bregman divergences for optimization and clustering problems [31–35].

The Bregman divergence includes many prominent dissimilarity measures like [26,23,36]:

- The generalized Kullback–Leibler (or I-) divergence for positive measures $p$ and $q$:

$$D_{GKL}(p\|q)=\int p\log\left(\frac{p}{q}\right)dr-\int(p-q)\,dr \tag{25}$$

using the generating function

$$\Phi(f)=\int(f\cdot\log f-f)\,dr. \tag{26}$$

Some three-dimensional isosurfaces for the generalized Kullback–Leibler divergence with respect to different reference points can be found in the first column of Figs. 2 and 4. Dependent on the choice of the divergence and its possible parameters the scaling and shape of the isosufaces vary. For probability densities $p$ and $q$, Eq. (25) simplifies to the Kullback–Leibler divergence [37,38]:

$$D_{KL}(p\|q)=\int p\log\left(\frac{p}{q}\right)dr, \tag{27}$$

which is related to the Shannon-entropy [39]. Equidistance contours for three-dimensional probability densities using Kullback–Leibler divergence with respect to different reference points are displayed in the first row of Figs. 3 and 5.

- The Itakura–Saito divergence [40]:

$$D_{IS}(p\|q)=\int\left[\frac{p}{q}-\log\left(\frac{p}{q}\right)-1\right]dr \tag{28}$$

bases on the Burg entropy, which also serves as the generating function:

$$\Phi(f)=-\int\log(f)\,dr. \tag{29}$$

The Itakura–Saito divergence was originally presented as a measure of the quality of fits between two spectra and became a standard measure in the speech and image processing community due to the good perceptual properties of the reconstructed signals. It is known as negative cross-Burg entropy and fulfills the scale-invariance property $D_{IS}(c\cdot p\|c\cdot q)=D_{IS}(p\|q)$, which implies the same relative weight is given to low and high components of $p$, see [41] for details.

- The Eta-divergence is also known as norm-like divergence [42]:

$$D_\eta(p\|q)=\int p^\eta+(\eta-1)\cdot q^\eta-\eta\cdot p\cdot q^{\eta-1}\,dr, \tag{30}$$

with generating function

$$\Phi(f)=\int f^\eta\,dr \quad \text{for } \eta>1. \tag{31}$$

In the case $\eta=2$ the Eta-divergence becomes the Euclidean distance with generating function $\Phi(f)=\int f^2\,dr$.

- The Beta-divergence [26]:

$$D_\beta(p\|q)=\int p\frac{p^{\beta-1}-q^{\beta-1}}{\beta-1}\,dr-\int\frac{p^\beta-q^\beta}{\beta}\,dr, \tag{32}$$

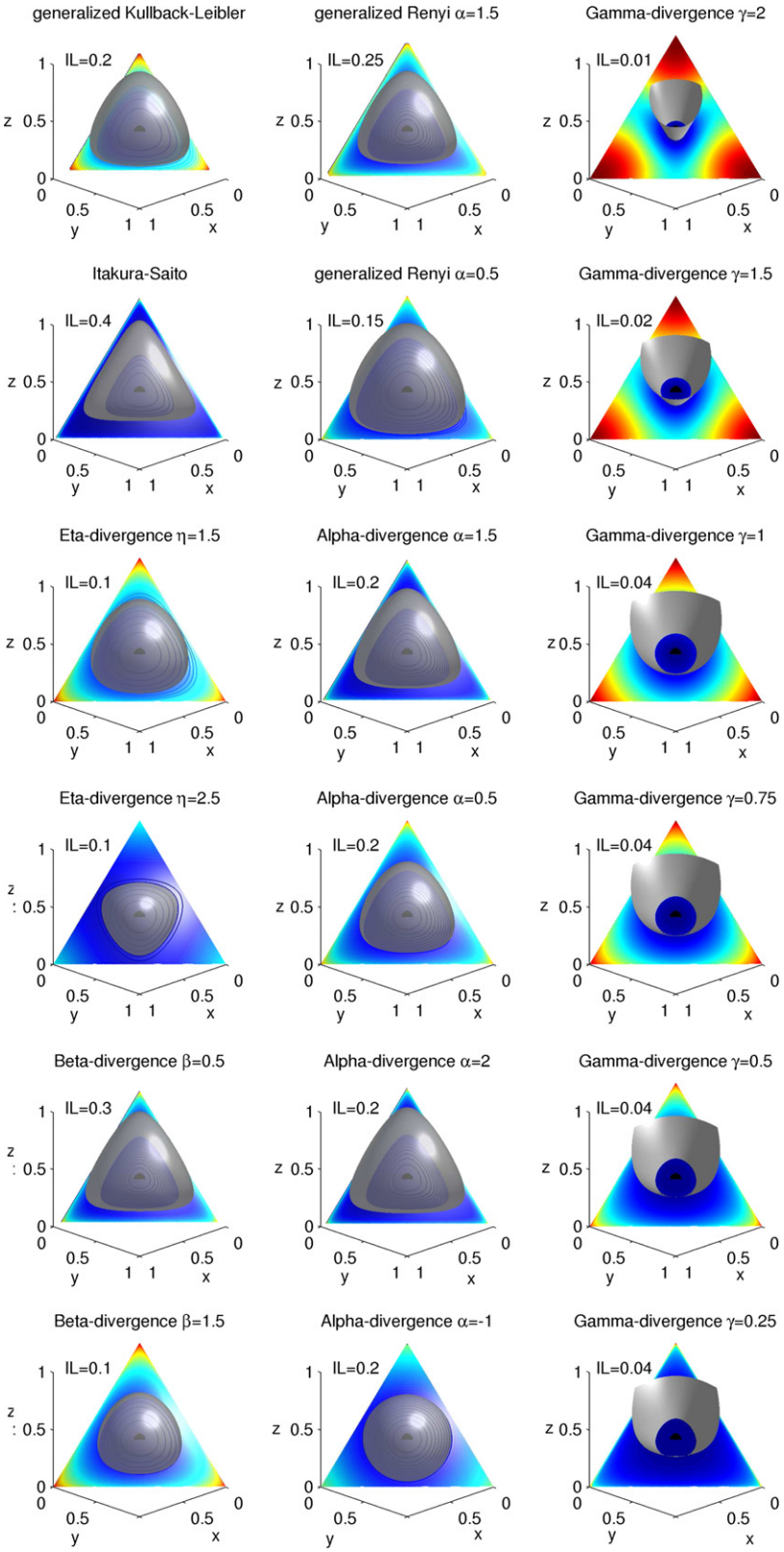


**Fig. 2.** Isosurfaces of some Example divergences including the plane of probability densities with respect to the reference point (0.3,0.3,0.3). The first column shows Bregman divergences, the second Csiszár-f divergences and the last column shows the Gamma divergence for different values of $\gamma$.

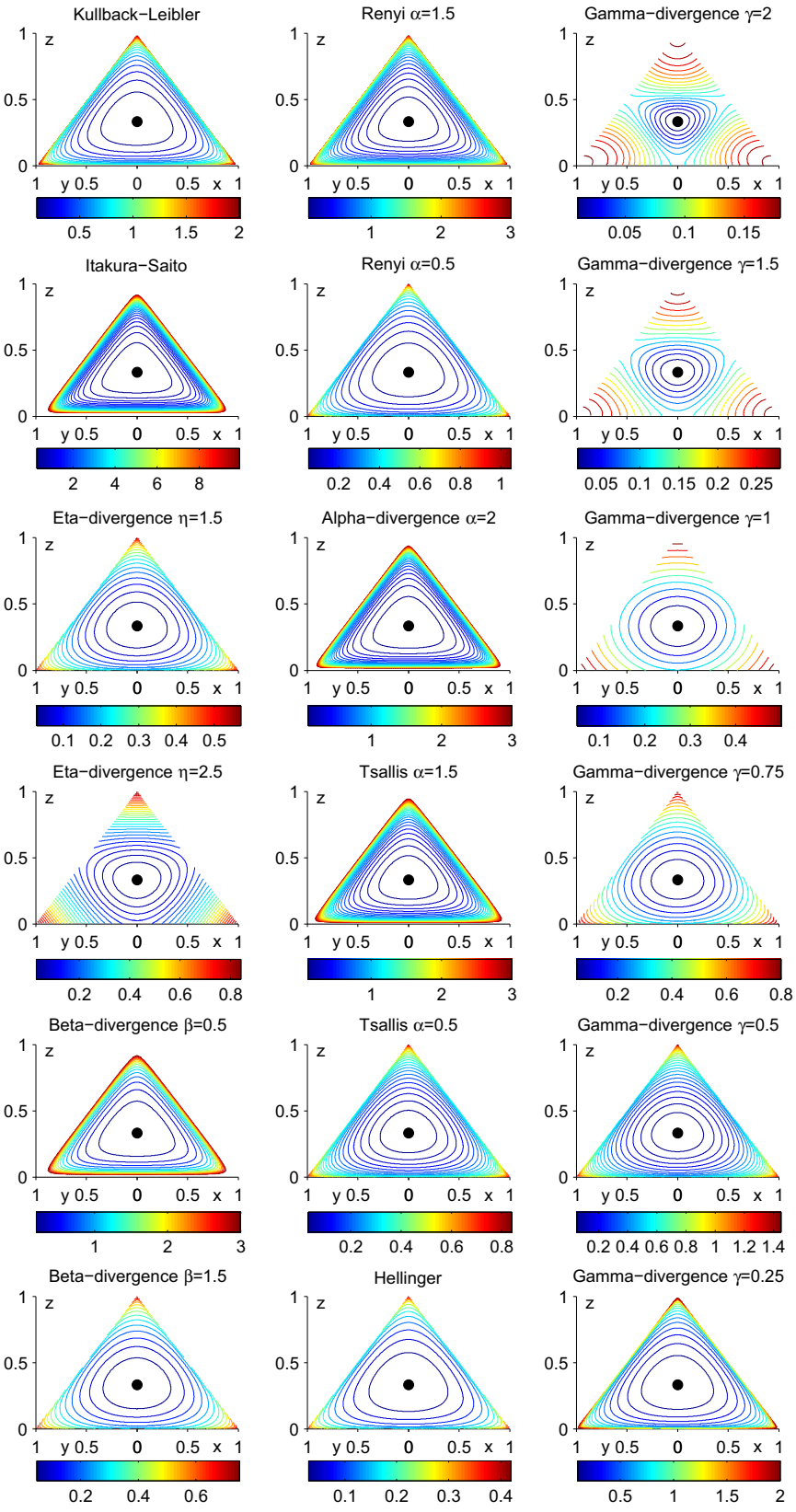


**Fig. 3.** Equidistance lines of some Example divergences for probability densities with respect to reference point (0.3,0.3,0.3). The columns show Bregman divergences, Csiszár-f divergences and Gamma divergences.

with $\beta \neq 0$ and $\beta \neq 1$ and the generating function

$$\Phi(f) = \frac{f^{\beta} - \beta \cdot f + \beta - 1}{\beta(\beta-1)}. \tag{33}$$

For specific values of $\beta$ the divergence becomes:
$\beta \to 1$: generalized Kullback–Leibler equation (25).
$\beta \to 0$: Itakura–Saito divergence equation (28).
$\beta = 2$: Euclidean distance (apart from a factor $\frac{1}{2}$).

Furthermore the Beta-divergence is equivalent to the density power divergence [43,36,44] and a rescaled version of the Eta-divergence.

### 4.2. Csiszár-f divergences

Csiszár-f divergences are connected with the "ratio test" in the Pearson–Neyman style hypothesis testing and are in many ways "natural" concerning distributions and statistics. We denote by $\mathcal{F}$ the class of convex, real-valued, continuous functions $f$ satisfying $f(1) = 0$, with

$$\mathcal{F} = \{g | g : [0,\infty) \to \mathbb{R}, g\text{–convex}\}. \tag{34}$$

For a function $f \in \mathcal{F}$ the Csizár $f$-divergence is given by

$$\mathrm{D}_f(p\|q) = \int q \cdot f\left(\frac{p}{q}\right) dr, \tag{35}$$

with the definitions $0 \cdot f(\frac{0}{0}) = 0$ and $0 \cdot f(a/0) = \lim_{r\to 0} r \cdot f(a/r) = \lim_{u\to\infty} a \cdot f(u)/u$ [45–48]. The $f$-divergence can be interpreted as an average of the likelihood ratio $p/q$ describing the change rate of $p$ with respect to $q$ weighted by the determining function $f$. For a general $f$, which does not have to be convex, with $f'(1) = c_f \neq 0$, this form is not invariant and we have to use the generalized $f$-divergence

$$\mathrm{D}_f^G(p\|q) = c_f \int (p-q)\, dr + \int q f\left(\frac{p}{q}\right) dr. \tag{36}$$

For the special case of probability densities $p$ and $q$ the first term vanishes and the original form of the $f$-divergences is obtained.

Some basic properties of the Csiszár $f$-divergence are [49,26]:

*Non-negativity*: $\mathrm{D}_f(p\|q) \geq 0$ where the equal sign holds iff $p \equiv q$, which follows from Jensen's inequality.

*Generalized entropy*: It corresponds to a generalized $f$-entropy if the form

$$H_f(p) = -\int f(p(r))\, dr. \tag{37}$$

*Strict convexity*: The $f$-divergence is convex in both arguments $p$ and $q$:

$$\mathrm{D}_f(tp_1 + (1-t)p_2 \| tq_1 + (1-t)q_2) \leq t\mathrm{D}_f(p_1\|q_1) + (1-t)\mathrm{D}_f(p_2\|q_2) \quad \forall t \in [0,1]. \tag{38}$$

*Scalability*: $c\mathrm{D}_f(p\|q) = \mathrm{D}_{cf}(p\|q)$ for any positive constant $c > 0$.

*Invariance*: $\mathrm{D}_f(p\|q)$ is invariant with respect to a linear shift regarding the function $f$: e.g. $\mathrm{D}_f(p\|q) = \mathrm{D}_{\tilde{f}}(p\|q)$ iff $\tilde{f}(u) = f(u) + c \cdot (u-1)$ for any constant $c \in \mathbb{R}$.

*Symmetry*: For $f, f^* \in \mathcal{F}$, where $f^*(u) = u \cdot f(1/u)$ denotes the conjugate function of $f$, the relation $\mathrm{D}_f(p\|q) = \mathrm{D}_{f^*}(q\|p)$ is valid. It is possible to construct a symmetric Csizár $f$-divergence with $f_{\mathrm{sym}}(u) = f(u) + f^*(u)$ as determining function.

*Upper bound*: The $f$-divergence is bounded by

$$0 \leq \mathrm{D}_f(p\|q) \leq \lim_{u\to 0^+} \{f(u) + f^*(u)\} \quad \text{with } u = \frac{p}{q}. \tag{39}$$

The existence of this limit for probability densities $p$ and $q$ was shown by Liese and Vajda in [50]. Villmann and Haase showed that these bounds still hold for positive measures $p$ and $q$ [23].

*Monotonicity*: The $f$-divergence is monotonic with respect to the coarse-graining of the underlying domain $\mathcal{D}$ of the positive measures $p$ and $q$, which is similar to the monotonicity of the Fisher metric [47].

Some well-known examples of $f$-divergences are:

- The subset of Alpha divergences [26]:

$$\mathrm{D}_\alpha(p\|q) = \frac{1}{\alpha(\alpha-1)} \cdot \int [p^\alpha q^{(1-\alpha)} - \alpha p + (\alpha-1)q]\, dr \tag{40}$$

  is based on the determining function

$$f(u) = u\frac{u^{(\alpha-1)} - 1}{\alpha^2 - \alpha} + \frac{1-u}{\alpha} \quad \text{with } u = \frac{p}{q}, \tag{41}$$

  with $\alpha \in \mathbb{R}\backslash\{0,1\}$. For specific values of $\alpha$ the divergence becomes [26]:
  $\alpha \to 1$: generalized Kullback–Leibler equation (25).
  $\alpha \to 0$: reverse Kullback–Leibler.
  $\alpha = -1$: Neyman Chi-square.
  $\alpha = 2$: Pearson Chi-square.

  For $\alpha \leq 0$ the divergence is zero-forcing, e.g. $p(r) = 0$ enforces $q(r) = 0$. On the other hand, for $\alpha \geq a$ it is zero-avoiding, i.e. $q(r) > 0$ whenever $p(r) > 0$. For $\alpha \to \infty$ $q(r)$ covers $p(r)$ completely and the Alpha divergence is called inclusive in this case. Furthermore the Beta-divergences can be generated from the Alpha divergences by applying a nonlinear transformation [26,23].
- The generalized Rényi divergence [51,26]:

$$\mathrm{D}_{\mathrm{GR}}^\alpha(p\|q) = \frac{1}{\alpha-1} \cdot \log\left(\int [p^\alpha q^{(1-\alpha)} - \alpha p + (\alpha-1)q]\, dr + 1\right) \tag{42}$$

  $\alpha \in \mathbb{R}\backslash\{0,1\}$ is closely related to the Alpha divergence.
- For the special case of probability densities the generalized Rényi divergence reduces to the Rényi divergence [52,53]:

$$\mathrm{D}_{\mathrm{R}}^\alpha(p\|q) = \frac{1}{\alpha-1} \log\left(\int p^\alpha q^{(1-\alpha)}\, dr\right), \tag{43}$$

  which bases on the Rényi entropy.
- The Tsallis-divergences:

$$\mathrm{D}_{\mathrm{T}}^\alpha(p\|q) = \frac{1}{1-\alpha}\left(1 - \int p^\alpha q^{(1-\alpha)}\, dr\right) \tag{44}$$

  for $\alpha \neq 1$ is a widely applied divergence for probability densities $p$ and $q$ based on the Tsallis entropy. It is also a rescaled version of the Alpha divergence. In the limit $\alpha \to 1$ it converges to the Kullback–Leibler divergence equation (27).
- The Hellinger divergence [48]:

$$\mathrm{D}_H(p\|q) = \frac{1}{2}\int (\sqrt{p} - \sqrt{q})^2\, dr, \tag{45}$$

  with generating function $f(u) = 2(1-\sqrt{u})$ for $u = p/q$ is defined for probability densities $p$ and $q$.

### 4.3. Gamma divergence

The Gamma divergence is very robust with respect to outliers [54] and was proposed by Fujisawa and Eguchi:

$$\mathrm{D}_\gamma(p\|q) = \log\left[\frac{\left[\int p^{\gamma+1}\, dr\right]^{1/(\gamma^2+\gamma)} \cdot \left[\int q^{\gamma+1}\, dr\right]^{1/(\gamma+1)}}{\left(\int p \cdot q^\gamma\, dr\right)^{1/\gamma}}\right]. \tag{46}$$

It is robust for $\gamma \in [0,1]$. In the limit $\gamma \to 0$ it becomes the Kullback–Leibler-divergence $\mathrm{D}_{\mathrm{KL}}(p\|q)$ for probability densities. For $\gamma = 1$ it becomes the Cauchy–Schwarz divergence:

$$\mathrm{D}_{\mathrm{CS}}(p\|q) = \frac{1}{2}\log\left(\int q^2\, dr \cdot \int p^2\, dr\right) - \log\left(\int p \cdot q\, dr\right), \tag{47}$$

which is based on the quadratic Rényi-entropy. The Cauchy–Schwarz divergence is symmetric and was introduced considering the Cauchy–Schwarz inequality for norms. It is frequently applied for Parzen window estimation, especially suitable for spectral clustering as well as related graph cut problems [55–57,23].

Some isosurfaces of the Gamma divergence for different values of $\gamma$ are shown in the last column of Figs. 2 and 4. The equidistance

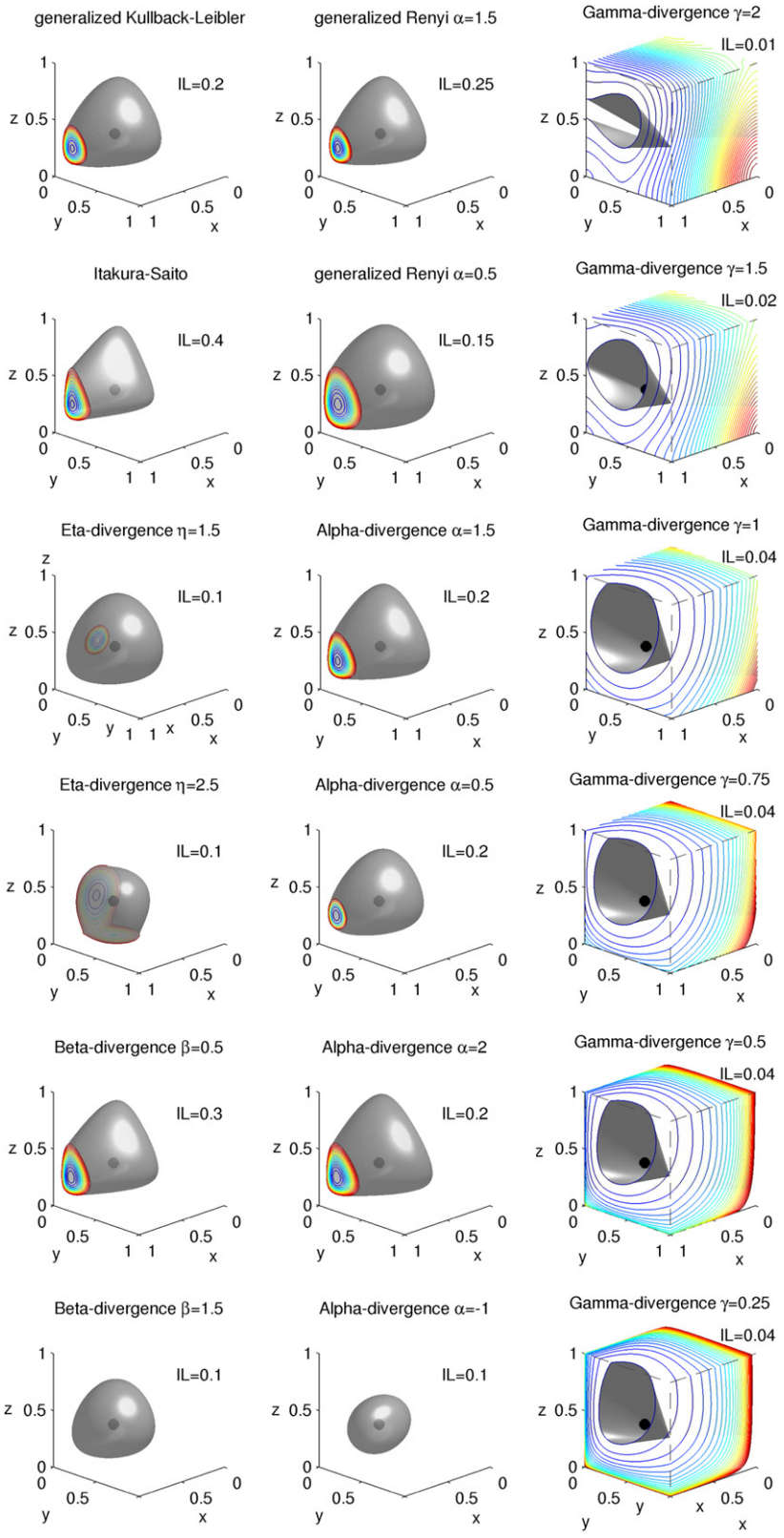


**Fig. 4.** Isosurfaces of some Example divergences with respect to the reference point (0.5,0.2,0.3). The cutoffs show the equidistance lines for this plane. The first column shows Bregman divergences, the second Csiszár-f divergences and the last column shows the Gamma divergence for different values of $\gamma$.

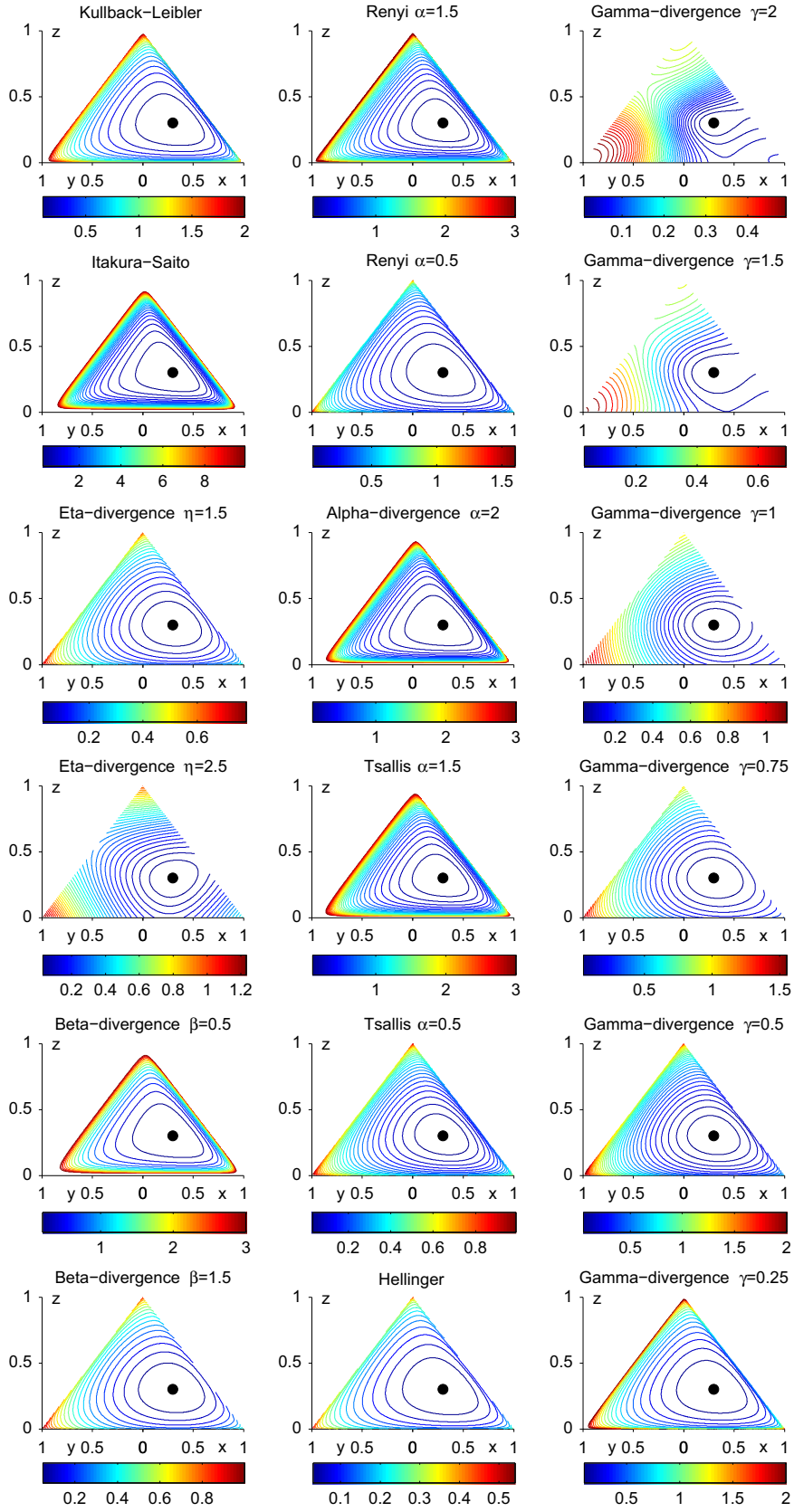


**Fig. 5.** Equidistance lines of some Example divergences for probability densities with respect to reference point (0.5,0.2,0.3). The columns show Bregman divergences, Csiszár-f divergences and Gamma divergences.

lines for the special case of probability densities can be found in the last column of Figs. 3 and 5. The Gamma divergence displays some nice properties [26,23]:

*Invariance*: $D_\gamma(p\|q)$ is invariant under scalar multiplication with positive constants

$$D_\gamma(p\|q) = D_\gamma(c_1 \cdot p\|c_2 \cdot q) \quad \forall c_1,c_2 > 0. \tag{48}$$

In case of positive measures the equation $D_\gamma(p\|q)=0$ holds only if $p = c \cdot q$ with $c>0$. For probability densities $c=1$ is required.

*Pythagorean relation*: As for Bregman divergences a modified Pythagorean relation between positive measures can be stated for special choices of $p,q,\rho$. Let $p$ be a distortion of $q$ defined as convex combination with a positive distortion measure $\phi(r)$

$$p_\varepsilon(r) = (1-\varepsilon)\cdot q(r) + \varepsilon \cdot \phi(r). \tag{49}$$

A positive measure $g$ is denoted as $\phi$-consistent if $v_g = (\int \phi(r)g(r)^\alpha\, dr)^{1/\alpha}$ is sufficiently small for large $\alpha>0$. If two positive measures $q$ and $\rho$ are $\phi$-consistent with respect to a distortion measure $\phi$, then the Pythagorean relation approximately holds for $q,\rho$ and the distortion $p_\varepsilon$ of $q$:

$$\Delta(p_\varepsilon,q,\rho) = D_\gamma(p_\varepsilon\|\rho) - D_\gamma(p_\varepsilon\|q) - D_\gamma(q\|\rho) = \mathcal{O}(\varepsilon v^\gamma) \quad \text{with } v = \max\{v_q, v_\rho\}. \tag{50}$$

This property implies the robustness of $D_\gamma$ according to distortions.

## 5. Discussion of divergences

In this section we examine and compare some introduced divergences by means of controlled experiments. We investigate the behavior of different divergences for the comparison of images containing an increasing level of (nonlinear) noise. Therefore, we compute the histograms of gray-value images taken from the Berkley segmentation data set and noisy versions of them.

### 5.1. Linearly monotonically increasing noise

In the first experiment the noisy image $I^*$ is obtained by adding a linear monotonically increasing transformation of gray values to the image $I$:

$$I^*(x,y) = I(x,y)\cdot[l\cdot(I(x,y)-I_0)+1], \tag{51}$$

where $l$ denotes the level of noise and $I_0$ corresponds to the minimal intensity in the original image. Fig. 6 shows a picture (in the following referred to as "moon") adding different levels of noise following Eq. (51) together with the gray-value histograms. The noise-level is ranged from $l=1$ to $l=9$. Some dissimilarity matrices comparing the 10 histograms with different divergence measures are shown in Fig. 7. The intuitively ideal dissimilarity matrix in this case is a symmetric band matrix shown in the middle of the top row. Some divergences like the generalized Rényi divergence show numerical instabilities. Others show quite similar behavior, e.g. Itakura–Saito, Alpha divergences and the Beta-divergence with $\beta=0.5$, but they do not exhibit the desired band structure. For the original image and low noise-levels (images 1–5) the Beta-divergence with $\beta=1.5$, Alpha divergence with $\alpha=0.5$ and also the generalized KL divergence show a bit of the desired band structure. Ignoring the last column and last row (the extreme case) in the dissimilarity matrix of the Eta-divergence shows a good approximation of a band matrix. The Gamma divergence is observed to be quite robust in this case and also exhibits a visible band structure for $\gamma \geq 1$. In the special case of $\gamma=1$ the Gamma divergence equals the Cauchy–Schwarz divergence and is symmetric. Another symmetric example is the Alpha divergence with $\alpha=0.5$.

As a second example we take a picture of a group of dolphins and add some noise (following Eq. (51)) using the levels $l=[0.1,0.2,\ldots,0.9]$. The resulting histograms of gray values for the different noise levels are shown in Fig. 8. As above we compute the matrices of pairwise similarities between the histograms using different divergences. The results can be found in Fig. 9. In this example the Eta-divergence especially with $\eta=2.5$ is a good approximation of the ideal dissimilarity matrix shown in the middle of the top row. The best symmetric choice is the Gamma divergence with $\gamma=1$ (Cauchy–Schwarz). Furthermore, dependent on the value for $\gamma$ one can chose between a better "resolution" (local) and a better preservation of the hierarchy of the histograms (global). Some other divergences, e.g. the generalized KL and Itakura–Saito, show very poor approximations of the desired dissimilarity for this example.

### 5.2. Additive uniform noise

In the second experiment the noisy image $I^*$ is obtained by adding uniform noise to the image $I$:

$$I^*(x,y) = I(x,y) + \mathcal{U}(0,l), \tag{52}$$

where $\mathcal{U}(0,l)$ denotes a scalar value drawn from the uniform distribution in the interval $[0,l]$.

Fig. 10 shows the picture of dolphins adding different levels of uniform noise following Eq. (52) together with the more and more flattened gray-value histograms. The noise-level is ranged from $l=\frac{50}{255}$ to $l=\frac{450}{255}$. Some dissimilarity matrices pairwise comparing the 10 images with different divergence measures are shown in Fig. 11. Some divergences like the generalized Rényi, Itakura–Saito and some Alpha- and Beta-divergences fail to approximate the desired band structure in the pairwise dissimilarity matrix.

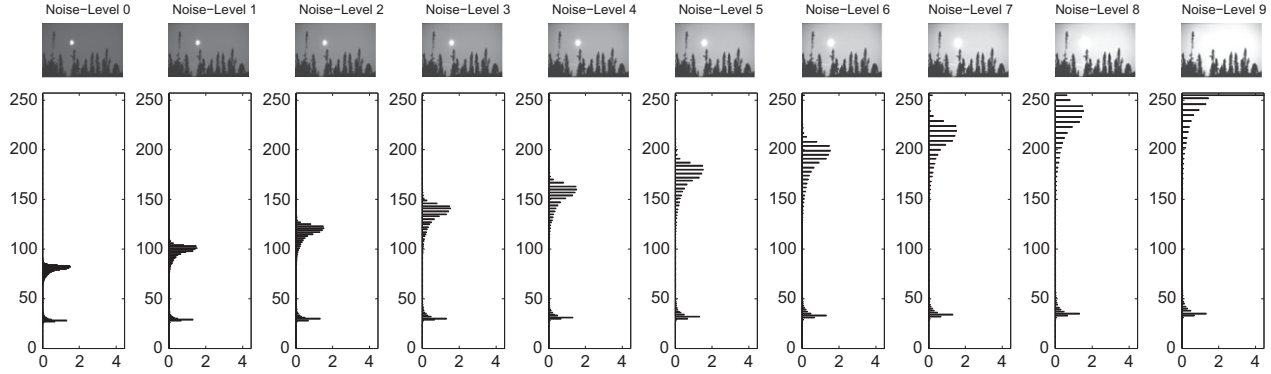


**Fig. 6.** Histograms of intensity values in an example picture. The original image "moon" (top row) together with its histogram is shown on the left side. The following pictures contain noise in the form of a linear monotonically increasing transformation of gray values following Eq. (51) using $l=[1,2,\ldots,9]$ corresponding to the Noise-Levels 1–9.

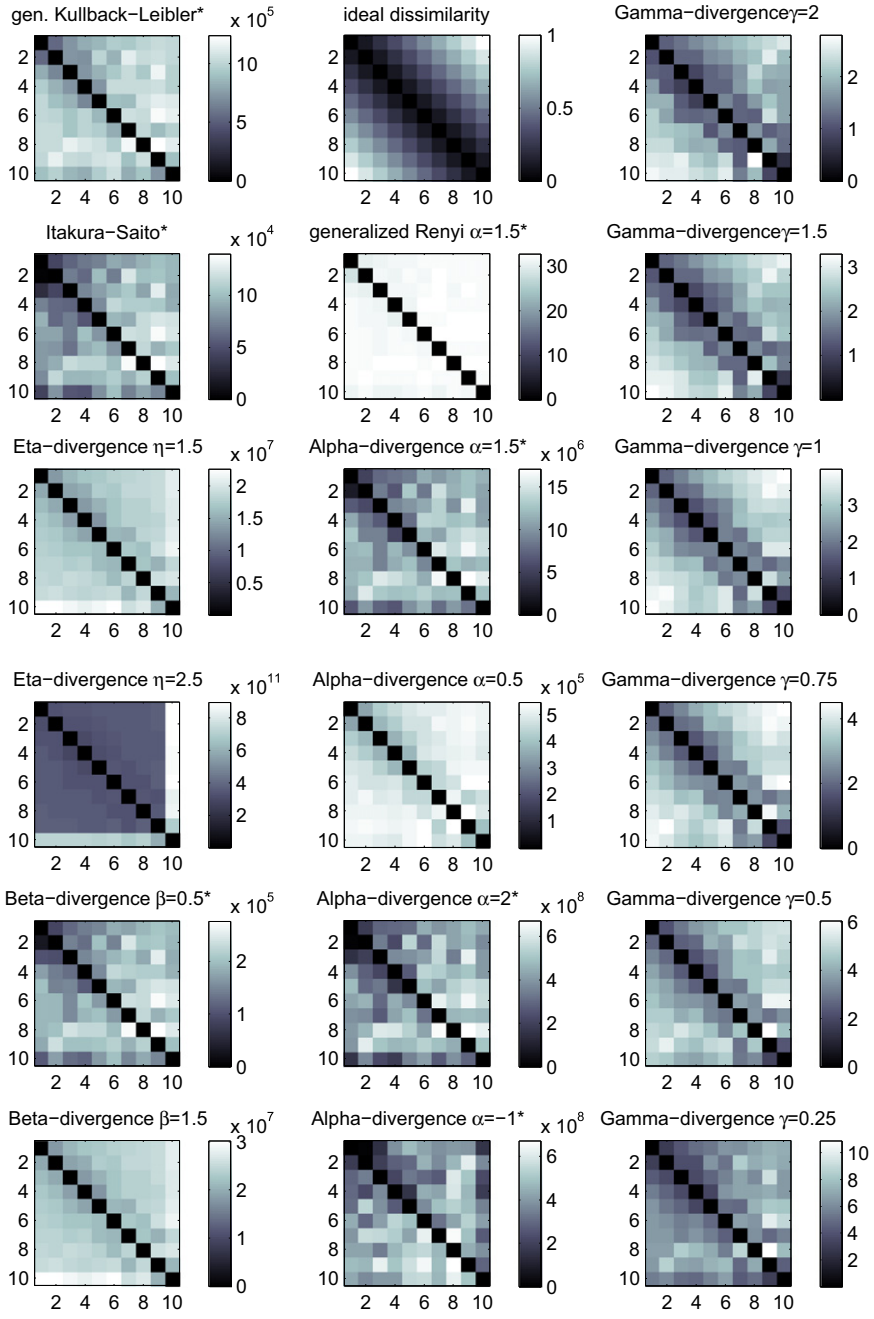


**Fig. 7.** Matrix of pairwise dissimilarity of the 10 histograms shown in Fig. 6 using different divergences. The ideal dissimilarity matrix for this example is a band matrix shown in the middle of the top row. Some divergences (marked with an asterisk ∗ in the title) show numerical instabilities in case of zeros in the signals. In that cases a small constant $c=1$ was added to all histograms to prevent the degeneration. Other divergences, like e.g. the Gamma divergence are more robust. The Eta-divergence ignoring the extreme cases and the Gamma divergence with $\gamma \geq 1$ exhibit more of the desired band structure for this example compared to other choices.

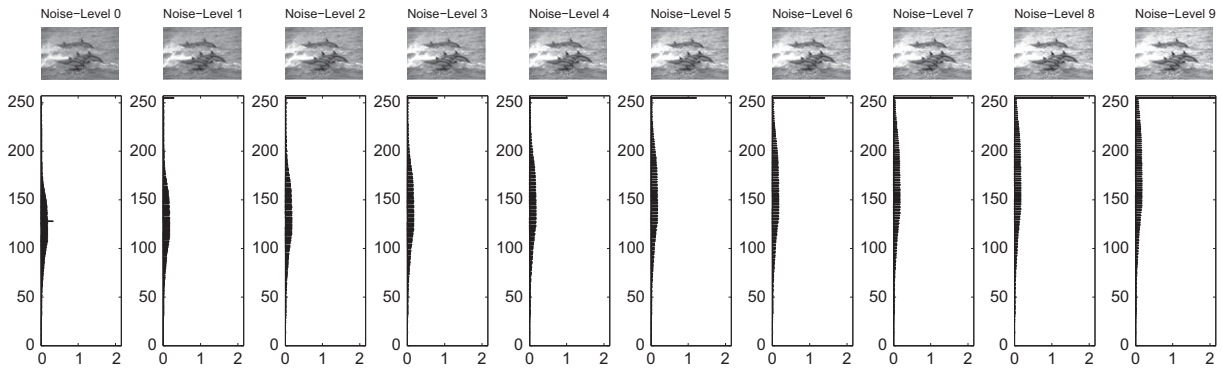


**Fig. 8.** Histograms of intensity values in an example picture. The original image "dolphins" (top row) together with its histogram is shown on the left side. The following pictures contain noise in the form of a linear monotonically increasing transformation of gray values following Eq. (51) using $l=[0.1,0.2,\ldots, 0.9]$ corresponding to the Noise-Levels 1–9.

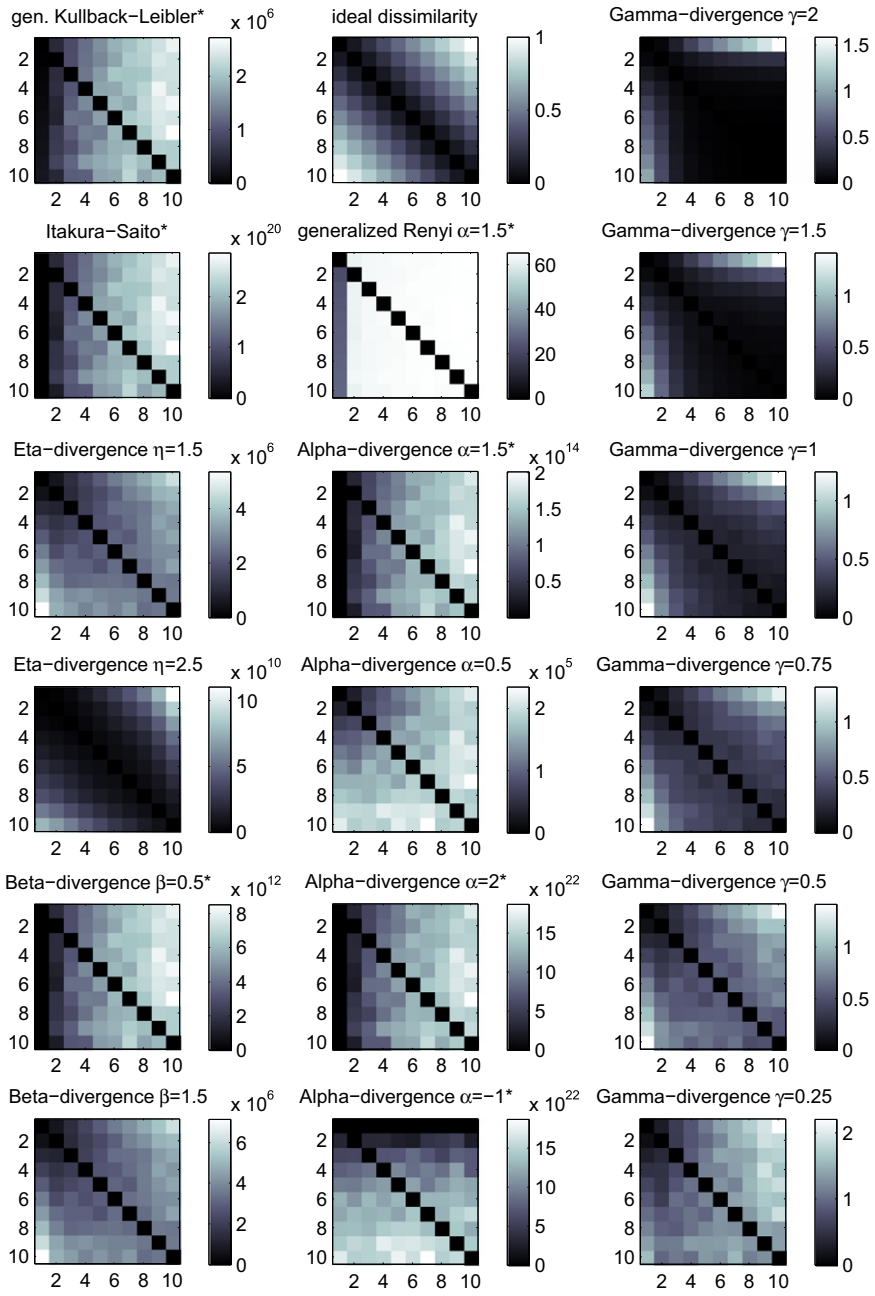


**Fig. 9.** Matrix of pairwise dissimilarity of the 10 histograms shown in Fig. 8 using different divergences. The ideal dissimilarity matrix for this example is a band matrix shown in the middle of the top row. Some divergences (marked with an asterisk ∗ in the title) show numerical instabilities in case of zeros in the signals. In that cases a small constant $c=1$ was added to all histograms to prevent the degeneration. The Eta-divergence especially with $\eta=2.5$ shows a good approximation of the desired band structure for this example. The Gamma divergence with $\gamma=1$ (Cauchy–Schwarz) is the best symmetric choice in this case.

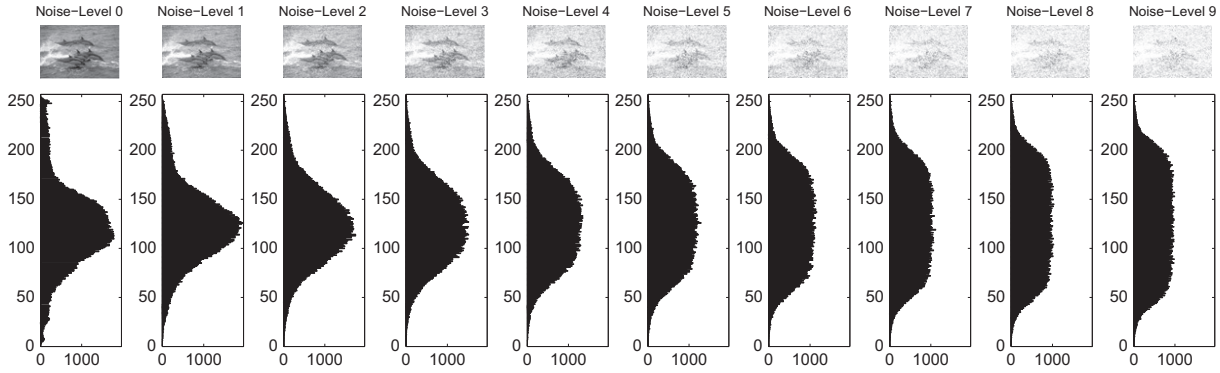


**Fig. 10.** Histograms of intensity values in an example picture. The original image "dolphins" (top row) together with its histogram is shown on the left side. The following pictures contain additive uniform noise following Eq. (52) using $l=[\frac{50}{255},\frac{100}{255},\ldots,\frac{450}{255}]$ corresponding to the Noise-Levels 1–9.

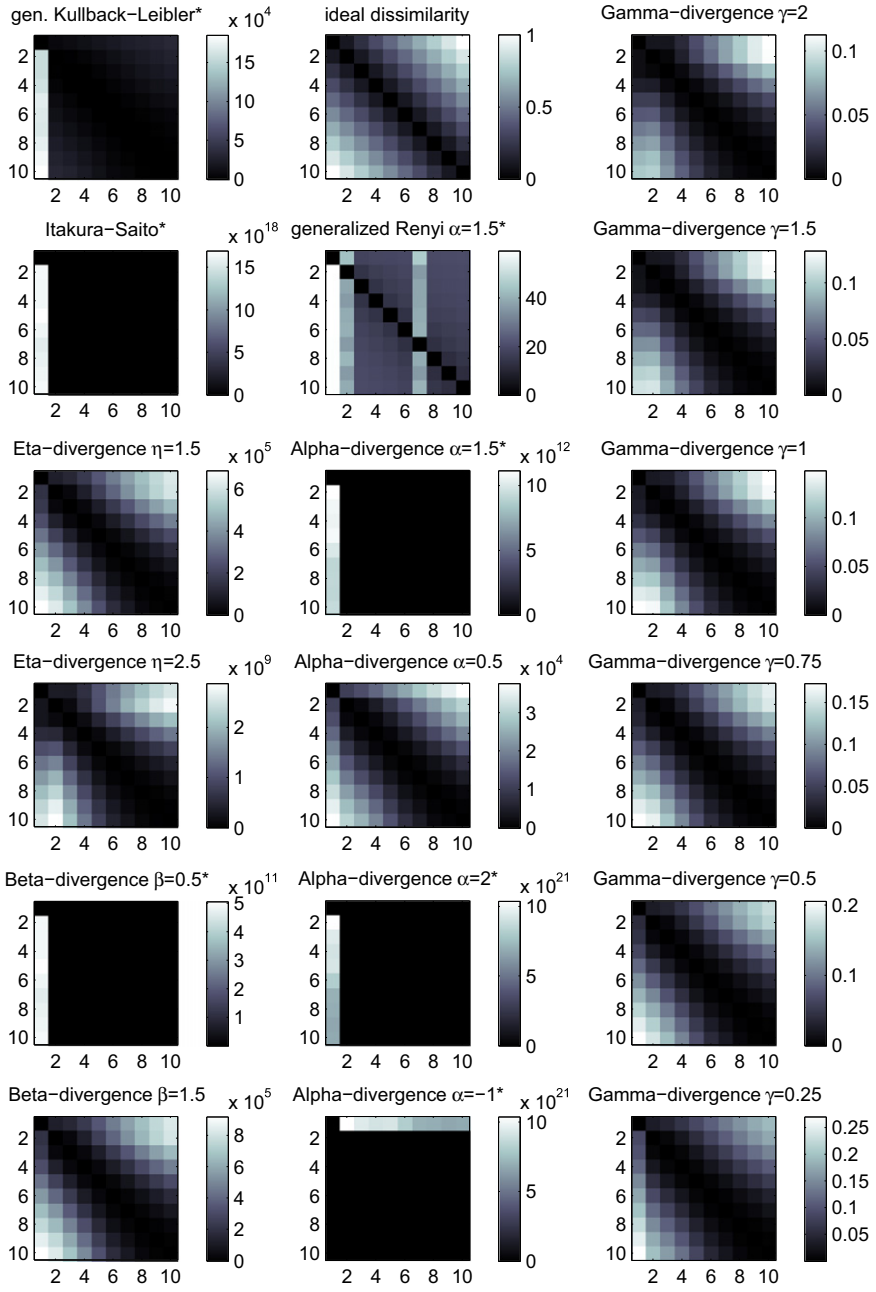


**Fig. 11.** Dissimilarity matrices comparing the 10 histograms shown in Fig. 10 using different divergences. The ideal dissimilarity matrix for this example is a band matrix shown in the middle of the top row. Some divergences (marked with an asterisk ∗ in the title) show numerical instabilities in case of zeros in the signals. In that cases a small constant $c=1$ was added to all histograms to prevent the degeneration. In this example the Eta-, Beta-, Gamma and the Alpha divergences with $\alpha=0.5$ show good approximations of the ideal band structure. Ignoring the original image also KL is nearly perfect. Other divergences like Itakura–Saito and generalized Rényi fail in this example.

Others, like the Gamma-, Eta- and some Alpha- and Beta-divergences are nearly ideal for this example. The Kullback–Leibler divergence is nearly perfect if the original image is ignored.

## 6. The Fréchet derivative

In this section we introduce the concept of Fréchet derivatives used for the generalization to arbitrary divergences. Suppose $V$ and $W$ are Banach spaces and $U\subset V$ is an open subset of $V$. The function $f: U\to W$ is called Fréchet differentiable at $r\in U$, if there exists a bounded linear operator $A_r: V\to W$, such that for $h\in U$

$$\lim_{h\to 0}\frac{\|f(r+h)-f(r)-A_r(h)\|_W}{\|h\|_V}=0. \tag{53}$$

This general definition can be used for functions $L: B\to\mathbb{R}$, defined as mappings from a functional Banach space $B$ to $\mathbb{R}$. Further let $B$ be equipped with a norm $\|\cdot\|$ and $f,h\in B$ are two functionals. The Fréchet derivative $\delta L[f]/\delta f$ of $L$ at point $f$ (i.e. in a function $f$) in the direction $h$ is formally defined as

$$\lim_{\epsilon\to 0}\frac{1}{\epsilon}(L[f+\epsilon h]-L[f]) =: \frac{\delta L[f]}{\delta f}[h]. \tag{54}$$

The Fréchet derivative in finite-dimensional spaces reduces to the usual partial derivative. Thus, it is a generalization of the directional derivatives.

Following [23] we introduce the functional derivatives of divergences in the next paragraphs. An overview is given in Table 3.

**Table 3**
Table of divergences and their Fréchet derivative.

| Divergence family | Formula | Fréchet derivative |
|---|---|---|
| Bregman divergence | $D_B^{\phi}(p\|q)=\phi(p)-\phi(q)-\frac{\delta\phi(q)}{\delta q}[p-q]$ | $\frac{\delta D_B^{\phi}(p\|q)}{\delta q}=\frac{\delta\phi(p)}{\delta q}-\frac{\delta\phi(q)}{\delta q}-\frac{\delta\left[\frac{\delta\phi(q)}{\delta q}(p-q)\right]}{\delta q}$ |
| Gen. Kullback–Leibler | $D_{GKL}(p\|q)=\int p\cdot\log\left(\frac{p}{q}\right)dr-\int(p-q)\,dr$ | $\frac{\delta D_{GKL}(p\|q)}{\delta q}=-\frac{p}{q}+1$ |
| Kullback–Leibler | $D_{KL}(p\|q)=\int p\cdot\log\left(\frac{p}{q}\right)dr$ | $\frac{\delta D_{KL}(p\|q)}{\delta q}=-\frac{p}{q}$ |
| Itakura–Saito | $D_{IS}(p\|q)=\int\left[\frac{p}{q}-\log\left(\frac{p}{q}\right)-1\right]dr$ | $\frac{\delta D_{IS}(p\|q)}{\delta q}=\frac{1}{q^2}(q-p)$ |
| Eta-divergence | $D_{\eta}(p\|q)=\int p^{\eta}+(\eta-1)\cdot q^{\eta}-\eta\cdot p\cdot q^{(\eta-1)}\,dr$ | $\frac{\delta D_{\eta}(p\|q)}{\delta q}=q^{(\eta-2)}\cdot(1-\eta)\cdot\eta\cdot(p-q)$ |
| Beta-divergence | $D_{\beta}(p\|q)=\int p\,\frac{p^{(\beta-1)}-q^{(\beta-1)}}{\beta-1}\,dr-\int\frac{p^{\beta}-q^{\beta}}{\beta}dr$ | $\frac{\delta D_{\beta}(p\|q)}{\delta q}=q^{(\beta-2)}(q-p)$ |
| Gen. Csiszár-f | $D_f^G(p\|q)=c_f\int(p-q)\,dr+\int qf\left(\frac{p}{q}\right)dr,\ c_f=f'(1)\neq 0$ | $\frac{\delta D_f^G(p\|q)}{\delta q}=f\left(\frac{p}{q}\right)q\frac{\partial f(u)}{\partial u}\cdot\frac{-p}{q^2},\ c_f=f'(1)\neq 0$ |
| Csiszár-f divergence | $D_f(p\|q)=\int q\cdot f\left(\frac{p}{q}\right)dr$ | $\frac{\delta D_f(p\|q)}{\delta q}=f\left(\frac{p}{q}\right)+q\frac{\partial f(u)}{\partial u}\cdot\frac{-p}{q^2}$ |
| Alpha divergence | $D_{\alpha}(p\|q)=\frac{1}{\alpha(\alpha-1)}\cdot\int[p^{\alpha}q^{(1-\alpha)}-\alpha p+(\alpha-1)q]\,dr$ | $\frac{\delta D_{\alpha}(p\|q)}{\delta q}=-\frac{1}{\alpha}(p^{\alpha}q^{(-\alpha)}-1)$ |
| Gen. Rényi | $D_{GR}^{\alpha}(p\|q)=\frac{1}{\alpha-1}\cdot\log\left(\int[p^{\alpha}q^{(1-\alpha)}-\alpha p+(\alpha-1)q]\,dr+1\right)$ | $\frac{\delta D_{GR}^{\alpha}(p\|q)}{\delta q}=\frac{-p^{\alpha}q^{(-\alpha)}-1}{\int[p^{\alpha}q^{(1-\alpha)}-\alpha\cdot p+(\alpha-1)q]\,dr+1}$ |
| Rényi | $D_R^{\alpha}(p\|q)=\frac{1}{\alpha-1}\cdot\log\left(\int p^{\alpha}q^{(1-\alpha)}\,dr\right)$ | $\frac{\delta D_R^{\alpha}(p\|q)}{\delta q}=\frac{-p^{\alpha}q^{(-\alpha)}}{\int p^{\alpha}q^{(1-\alpha)}\,dr}$ |
| Tsallis | $D_T^{\alpha}(p\|q)=\frac{1}{1-\alpha}\left(1-\int p^{\alpha}q^{(1-\alpha)}\,dr\right)$ | $\frac{\delta D_T^{\alpha}(p\|q)}{\delta q}=\frac{-p^{\alpha}q^{(-\alpha)}}{\int p^{\alpha}q^{(1-\alpha)}\,dr}$ |
| Hellinger | $D_H(p\|q)=\frac{1}{2}\int(\sqrt{p}-\sqrt{q})^2\,dr$ | $\frac{\delta D_H(p\|q)}{\delta q}=1-\sqrt{\frac{p}{q}}$ |
| Gamma | $D_{\gamma}(p\|q)=\log\left[\frac{(\int p^{(\gamma+1)}\,dr)^{1/\gamma(\gamma+1)}\cdot(\int q^{(\gamma+1)}\,dr)^{1/(\gamma+1)}}{(\int p\cdot q^{\gamma}\,dr)^{1/\gamma}}\right]$ | $\frac{\delta D_{\gamma}(p\|q)}{\delta q}=\frac{q^{\gamma}}{\int q^{(\gamma+1)}\,dr}-\frac{p\cdot q^{(\gamma-1)}}{\int p\cdot q^{\gamma}\,dr}$ |
| Cauchy– Schwarz | $D_{CS}(p\|q)=\frac{1}{2}\cdot\log(\int q^2\,dr\cdot\int p^2\,dr)-\log(\int p\cdot q\,dr)$ | $\frac{\delta D_{CS}(p\|q)}{\delta q}=\frac{q}{\int q^2\,dr}-\frac{p}{\int p\cdot q\,dr}$ |

### 6.1. Fréchet derivatives: Bregman divergences

The Fréchet-derivative of $D_B^{\phi}$ Eq. (19) with respect to $q$ is formally given by

$$\frac{\delta D_B^{\phi}(p\|q)}{\delta q}=\frac{\delta\phi(p)}{\delta q}-\frac{\delta\phi(q)}{\delta q}-\frac{\delta\left[\frac{\delta\phi(q)}{\delta q}(p-q)\right]}{\delta q},$$

with

$$\frac{\delta\left[\frac{\delta\phi(q)}{\delta q}(p-q)\right]}{\delta q}=\frac{\delta^2[\phi(q)]}{\delta q^2}(p-q)-\frac{\delta\phi(q)}{\delta q}.$$

For the generalized Kullback–Leibler divergence equation (25) this simplifies to

$$\frac{\delta D_{GKL}(p\|q)}{\delta q}=-\frac{p}{q}+1, \tag{55}$$

whereas for the Kullback–Leibler divergence equation (27) in the special case of probability densities it reads

$$\frac{\delta D_{KL}(p\|q)}{\delta q}=-\frac{p}{q}. \tag{56}$$

For the Itakura–Saito divergence equation (28) we get

$$\frac{\delta D_{IS}(p\|q)}{\delta q}=\frac{1}{q^2}(q-p) \tag{57}$$

and for the Eta-divergence equation (30) the Fréchet-derivative is

$$\frac{\delta D_{\eta}(p\|q)}{\delta q}=q^{(\eta-2)}\cdot(1-\eta)\cdot\eta\cdot(p-q). \tag{58}$$

In the case of $\eta=2$ it reduces to the derivative of the Euclidean distance $-2(p-q)$. The Fréchet-derivative for the subset of Beta-divergences equation (32) is given by

$$\frac{\delta D_{\beta}(p\|q)}{\delta q}=-p\cdot q^{(\beta-2)}+q^{(\beta-1)}, \tag{59}$$

$$\frac{\delta D_{\beta}(p\|q)}{\delta q}=q^{(\beta-2)}(q-p). \tag{60}$$

### 6.2. Fréchet derivatives: Csiszár-f divergences

For the Csiszár-f divergences equation (35) the Fréchet derivative is

$$\frac{\delta D_f(p\|q)}{\delta q}=f\left(\frac{p}{q}\right)+q\frac{\partial f(u)}{\partial u}\frac{\delta u}{\delta q}, \tag{61}$$

$$\frac{\delta D_f(p\|q)}{\delta q}=f\left(\frac{p}{q}\right)+q\frac{\partial f(u)}{\partial u}\cdot\frac{-p}{q^2}, \tag{62}$$

with $u=p/q$. For the set of Alpha divergences equation (40) we get

$$\frac{\delta D_{\alpha}(p\|q)}{\delta q}=-\frac{1}{\alpha}(p^{\alpha}q^{(-\alpha)}-1). \tag{63}$$

The related generalized Rényi divergence equation (42) yields

$$\frac{\delta D_{GR}^{\alpha}(p\|q)}{\delta q}=\frac{-p^{\alpha}q^{(-\alpha)}-1}{\int[p^{\alpha}q^{(1-\alpha)}-\alpha p+(\alpha-1)q]\,dr+1}, \tag{64}$$

which reduces in the case of the Rényi divergence for probability densities to

$$\frac{\delta \mathrm{D}_{\mathrm{R}}^{\alpha}(p\|q)}{\delta q} = \frac{-p^{\alpha}q^{(-\alpha)}}{\int p^{\alpha}q^{(1-\alpha)}\,dr}. \tag{65}$$

For the Tsallis divergence equation (44) the Fréchet derivative reads

$$\frac{\delta \mathrm{D}_{\mathrm{T}}^{\alpha}(p\|q)}{\delta q} = \frac{-p^{\alpha}q^{(-\alpha)}}{\int p^{\alpha}q^{(1-\alpha)}\,dr} \tag{66}$$

and for the well-known Hellinger divergence equation (45) the derivative is

$$\frac{\delta \mathrm{D}_{\mathrm{H}}(p\|q)}{\delta q} = 1-\sqrt{\frac{p}{q}}. \tag{67}$$

### 6.3. Fréchet derivatives: Gamma divergences

The Fréchet derivative of the Gamma divergence equation (46) can be written as

$$\frac{\delta \mathrm{D}_{\gamma}(p\|q)}{\delta q} = \frac{q^{\gamma}}{\int q^{(\gamma+1)}\,dr} - \frac{p\cdot q^{(\gamma-1)}}{\int p\cdot q^{\gamma}\,dr}. \tag{68}$$

Considering the important special case $\gamma=1$, i.e. Cauchy–Schwarz divergence equation (47),

$$\frac{\delta \mathrm{D}_{\mathrm{CS}}(p\|q)}{\delta q} = \frac{q}{\int q^2\,dr} - \frac{p}{\int p\cdot q\,dr}. \tag{69}$$

## 7. t-SNE gradients for various divergences

In this section we explain the t-SNE gradients for various divergences. There exists a large variety of divergences which can be collected into several classes according to their mathematical properties and structural behavior. Here we follow the classification proposed in [26]. For this purpose, we plug the corresponding Fréchet-derivatives into the general gradient equation (14) for t-SNE. Clearly, one can convey these results easily to the general SNE gradient equation (16) in complete analogy, because of its structural similarity to the t-SNE formula equation (14).

A technical remark should be made here: In the following we will abbreviate $p(r)$ by $p$ and $p(r')$ by $p'$. Further, because the integration variable $r$ is a function $r=r(\xi,\zeta)$ an integration requires the weighting according to the distribution $\Pi_r$. Thus, the integration has formally to be carried out according to the differential $d\Pi_r(r)$ (Stieltjes-integral). We abbreviate this by $dr$ but keeping this fact in mind, i.e. by this convention, we will drop the distribution $\Pi_r$, if it is clear from the context.

### 7.1. Bregman divergences

In the following we will provide the Gradients for some examples of Bregman divergences introduced in Section 4.1. As a first example we show that we obtain the same result as van der Maaten and Hinton in [16] for the Kullback–Leibler divergence equation (27). The Fréchet-derivative of $D_{\mathrm{KL}}$ with respect to $q$ is given in Eq. (56). From Eq. (14) we see that

$$\frac{\partial \mathrm{D}_{\mathrm{KL}}}{\partial \xi} = 4\int \frac{q(\xi-\zeta)}{(1+r)}\left(\frac{p}{q} - \int p'\Pi_{r'}\,dr'\right)d\zeta. \tag{70}$$

Since the Integral $I=\int p'\Pi_{r'}\,dr'$ in Eq. (70) can be written as an double integral over all pairs of data points $I=\iint p'\,d\xi'\,d\zeta'$, we see from Eq. (8) that the integral $I$ equals 1. So, Eq. (70) simplifies to

$$\frac{\partial \mathrm{D}_{\mathrm{KL}}}{\partial \xi} = 4\int (1+r)^{-1}(p-q)(\xi-\zeta)\,d\zeta.$$

This is exactly the differential form of the discrete version as proposed for t-SNE in [16].

The Kullback–Leibler divergence used in original SNE and t-SNE belongs to the more general class of Bregman divergences [32]. Another representative of this class of divergences is the Itakura–Saito divergence $D_{\mathrm{IS}}$ equation (28) with the Fréchet-derivative equation (57). For the calculation of the gradient $\partial \mathrm{D}_{\mathrm{IS}}/\partial\xi$ we substitute the Fréchet-derivative in Eq. (14) and obtain

$$\begin{aligned}\frac{\partial \mathrm{D}_{\mathrm{IS}}}{\partial \xi} &= -4\int \frac{q}{1+r}\left(\frac{1}{q^2}(q-p) - \int \frac{q'-p'}{q'}\Pi_{r'}\,dr'\right)(\xi-\zeta)\,d\zeta \\ &= \int \frac{4(\xi-\zeta)}{1+r}\left[\frac{p}{q} - 1 + q\int\left[1-\frac{p'}{q'}\right]\Pi_{r'}\,dr'\right]d\zeta.\end{aligned} \tag{71}$$

One more Bregman-divergence is the norm-like or Eta-divergence equation (30). The Fréchet-derivative of $D_{\eta}$ with respect to $q$ is given in Eq. (58). Again, we are interested in the gradient $\partial \mathrm{D}_{\eta}/\partial\xi$, which is

$$\frac{\partial \mathrm{D}_{\eta}}{\partial \xi} = 4\eta(\eta-1)\int \frac{\xi-\zeta}{1+r}\left((p-q)q^{\eta-1} - q\cdot\int (p'-q')q'^{(\eta-1)}\Pi_{r'}\,dr'\right)d\zeta. \tag{72}$$

The last example of Bregman-divergences we handle in this paper is the class of Beta-divergences defined in Eq. (32). We use Eq. (14) and insert the Fréchet-derivative of the Beta-divergences, given by Eq. (59). Thereby the gradient $\partial \mathrm{D}_{\beta}/\partial\xi$ reads as

$$\frac{\partial \mathrm{D}_{\beta}}{\partial \xi} = 4\int \frac{\xi-\zeta}{1+r}\left(q^{\beta-1}(p-q) - q\cdot\int q'^{(\beta-1)}(p'-q')\Pi_{r'}\,dr'\right)d\zeta. \tag{73}$$

### 7.2. Csiszár's f-divergences

Now we will consider some divergences belonging to the class of Csiszár's $f$-divergences (see Section 4.2). A well-known example is the Hellinger divergence defined in Eq. (45), with the Fréchet-derivative equation (67). The gradient of $\mathrm{D}_{\mathrm{H}}$ with respect to $\xi$ is

$$\begin{aligned}\frac{\partial \mathrm{D}_{\mathrm{H}}}{\partial \xi} &= 4\int \frac{1}{1+r}\left(\sqrt{pq} - q - q\int(\sqrt{p'q'} - q')\Pi_{r'}\,dr'\right)(\xi-\zeta)\,d\zeta \\ &= 4\int \frac{\xi-\zeta}{1+r}\left[\sqrt{pq} - q\int \sqrt{p'q'}\Pi_{r'}\,dr'\right]d\zeta.\end{aligned} \tag{74}$$

For the Alpha divergence, see Eqs. (40) and (63), we get

$$\begin{aligned}\frac{\partial \mathrm{D}_{\alpha}}{\partial \xi} &= \frac{4}{\alpha}\int \frac{q(\xi-\zeta)}{1+r}\left(p^{\alpha}q^{(-\alpha)} - 1 - \int (p'^{\alpha}q'^{(-\alpha)} - 1)q'\Pi_{r'}\,dr'\right)d\zeta \\ &= \frac{4}{\alpha}\int \frac{\xi-\zeta}{1+r}\left[p^{\alpha}q^{(1-\alpha)} - q\int p'^{\alpha}q'^{(1-\alpha)}\Pi_{r'}\,dr'\right]d\zeta.\end{aligned} \tag{75}$$

For the Tsallis divergence, Eqs. (44) and (66), we get

$$\begin{aligned}\frac{\partial \mathrm{D}_{\alpha}^{\mathrm{T}}}{\partial \xi} &= \int \frac{4(\xi-\zeta)q}{1+r}\left[\left[\frac{p}{q}\right]^{\alpha} - \int \left[\frac{p'}{q'}\right]^{\alpha} q'\Pi_{r'}\,dr'\right]d\zeta \\ &= 4\int \frac{\xi-\zeta}{1+r}\left[p^{\alpha}q^{(1-\alpha)} - q\int p'^{\alpha}q'^{(1-\alpha)}\Pi_{r'}\,dr'\right]d\zeta,\end{aligned} \tag{76}$$

which is also clear from Eq. (75), since the Tsallis-divergence is a rescaled version of the Alpha divergence for probability densities.

For the Rényi-divergences, Eqs. (43) and (65), the derivative reads

$$\begin{aligned}\frac{\partial \mathrm{D}_{\mathrm{R}}^{\alpha}}{\partial \xi} &= \frac{4}{\int p'^{\alpha}q'^{(1-\alpha)}\,dr'}\int \frac{\xi-\zeta}{1+r}\left(p^{\alpha}q^{1-\alpha} - q\int p'^{\alpha}q'^{(1-\alpha)}\Pi_{r'}\,dr'\right)d\zeta \\ &= 4\int \frac{\xi-\zeta}{1+r}\left(\frac{p^{\alpha}q^{(1-\alpha)}}{\int p'^{\alpha}q'^{(1-\alpha)}\,dr'} - q\right)d\zeta.\end{aligned} \tag{77}$$

### 7.3. Gamma divergences

The Fréchet-derivative of $D_\gamma(p\|q)$ with respect to $q$ is given in Eq. (46) can be rewritten as

$$\frac{\delta D_\gamma(p\|q)}{\delta q} = q^{(\gamma-1)}\left[\frac{q}{\int q^{(\gamma+1)}\,dr} - \frac{p}{\int pq^\gamma\,dr}\right] = \frac{q^\gamma}{Q_\gamma} - \frac{pq^{(\gamma-1)}}{V_\gamma}$$

$$= \frac{q^\gamma V_\gamma - pq^{(\gamma-1)}Q_\gamma}{Q_\gamma V_\gamma}.$$

Once again, we use Eq. (14) to calculate the gradient of $D_\gamma$ with respect to $\xi$:

$$\frac{\partial D_\gamma}{\partial \xi} = \frac{-4}{Q_\gamma V_\gamma}\int \frac{q(\xi-\zeta)}{1+r}\Big[q^\gamma V_\gamma - pq^{(\gamma-1)}Q_\gamma - \int (q'^\gamma V_\gamma - p'q'^{(\gamma-1)}Q_\gamma)q'\Pi_{r'}\,dr'\Big]d\zeta$$

$$= -\frac{4}{Q_\gamma V_\gamma}\int \frac{q(\xi-\zeta)}{1+r}\Big(q^\gamma V_\gamma - pq^{(\gamma-1)}Q_\gamma - V_\gamma \times \int q'^{(\gamma+1)}\Pi_{r'}\,dr' + Q_\gamma\int p'q'^\gamma\Pi_{r'}\,dr'\Big)d\zeta$$

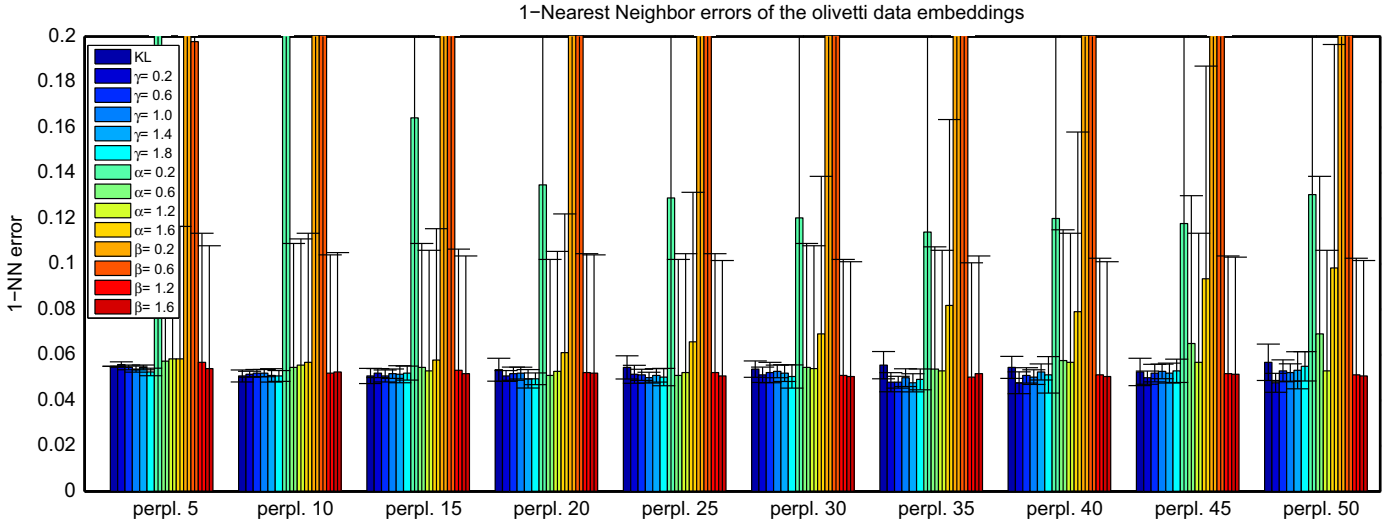


**Fig. 12.** Nearest neighbor errors of the two-dimensional embeddings using the Gamma-, Renyi- and Beta-divergence on the Olivetti faces data in comparison with the original formulation using Kullback–Leibler (KL) for different perplexities.

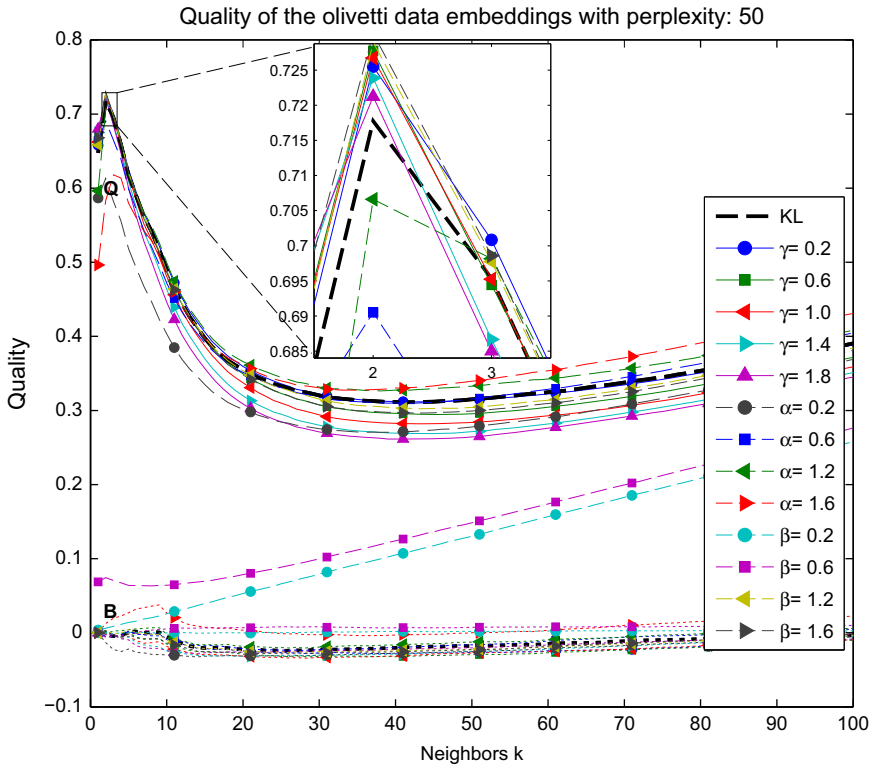


**Fig. 13.** Quality of the two-dimensional embeddings using the Gamma-, Renyi- and Beta-divergence on the Olivetti faces data in comparison with the original formulation using Kullback–Leibler (KL).

$$= -\frac{4}{Q_\gamma V_\gamma}\int \frac{q(\xi-\zeta)}{1+r}(q^\gamma V_\gamma - pq^{\gamma-1}Q_\gamma - V_\gamma Q_\gamma + Q_\gamma V_\gamma)d\zeta$$

$$= 4\int \frac{\xi-\zeta}{1+r}\left(\frac{pq^\gamma}{\int p'q'^\gamma\, dr'} - \frac{q^{(\gamma+1)}}{\int q'^{(\gamma+1)}\, dr'}\right)d\zeta. \quad (78)$$

For the special choice $\gamma = 1$ the Gamma divergence becomes the Cauchy–Schwarz divergence equation (47) and the gradient $\partial D_{CS}/\partial\xi$ for t-SNE can be directly derived from Eq. (78):

$$\frac{\partial D_{CS}}{\partial \xi} = 4\int \frac{\xi-\zeta}{1+r}\left(\frac{pq}{\int p'q'\, dr'} - \frac{q^2}{\int q'^2\, dr'}\right)d\zeta. \quad (79)$$

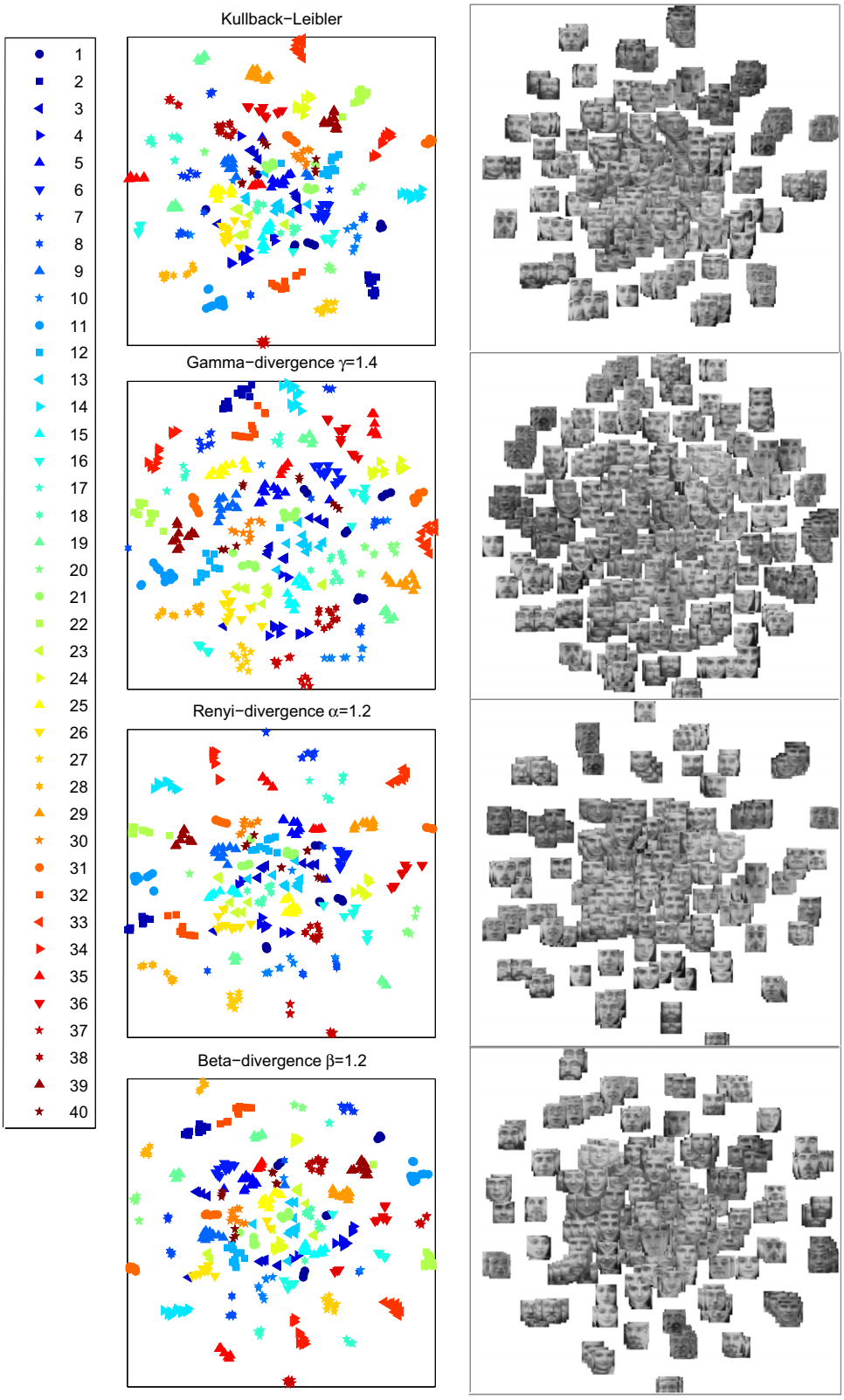


**Fig. 14.** Embeddings of the Olivetti faces based on the same initialization for different divergences and perplexity 20.

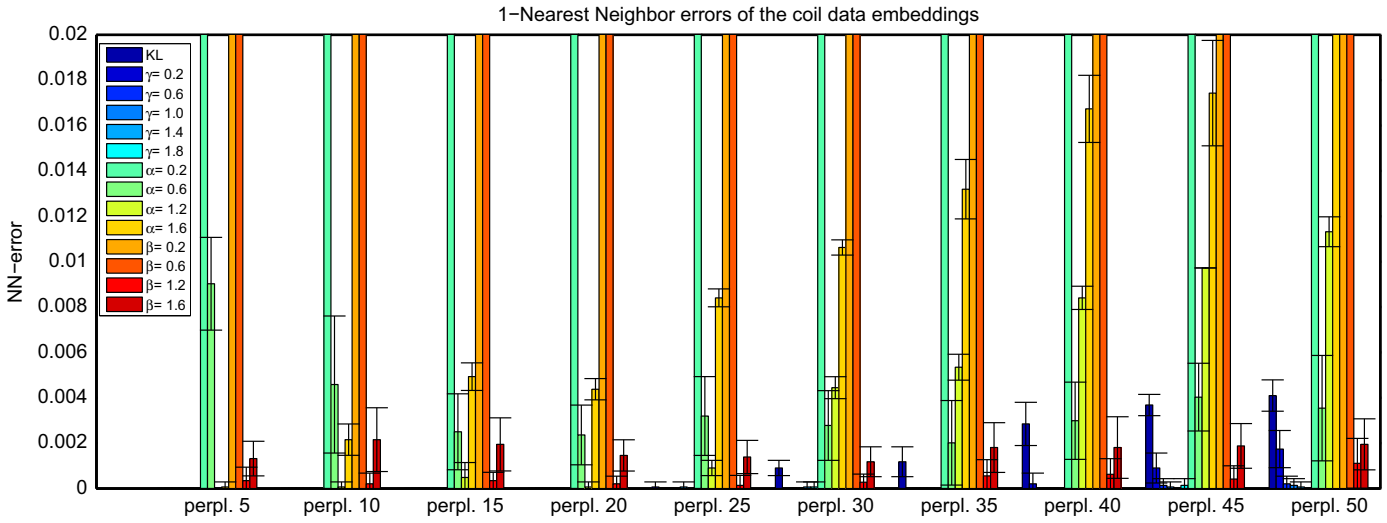


**Fig. 15.** Nearest neighbor errors of the two-dimensional embeddings using the Gamma-, Renyi- and Beta-divergence on the COIL-20 data in comparison with the original formulation using Kullback–Leibler (KL) for different perplexities.

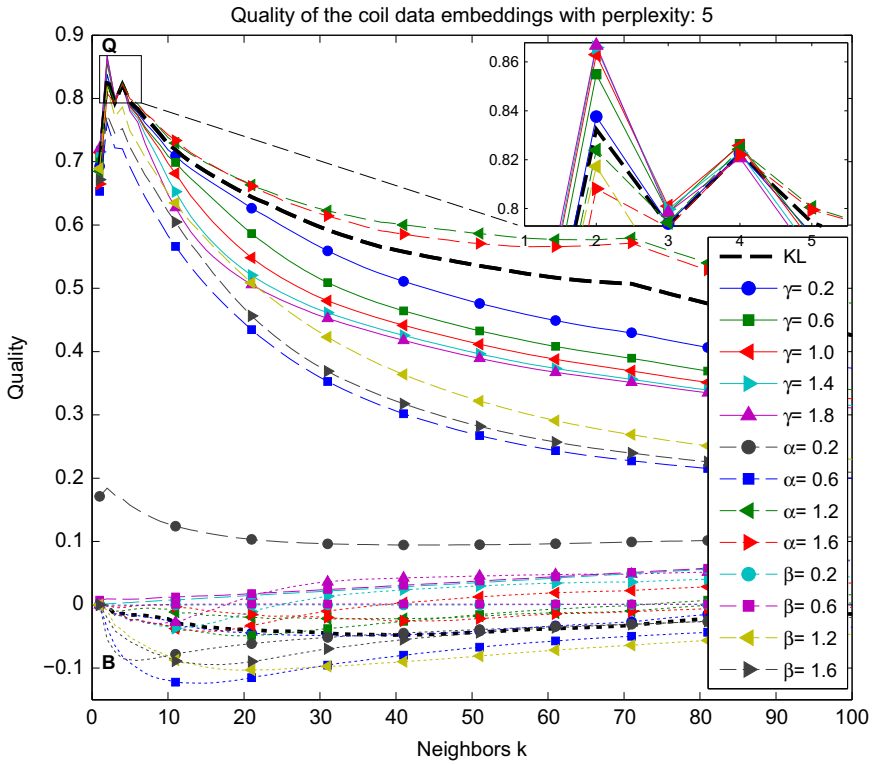


**Fig. 16.** Quality of the two-dimensional embeddings using the Gamma-, Renyi- and Beta-divergence on the COIL-20 data in comparison with the original formulation using Kullback–Leibler (KL).

Moreover, similar derivations can be made for any other divergence, since one only needs to calculate the Fréchet-derivative of the divergence and apply it to Eq. (14).

## 8. Demonstration of different divergences

In this section we demonstrate the use of different divergences in the t-SNE method on the bases of the Olivetti faces data set[1] and the COIL-20 data set [58]. In the experiments we compare one divergence from all three main families: Kullback–Leibler, Beta, Rényi and Gamma as example for Bregman-, Csiszár-f- and Gamma divergences. For the Gamma divergence we include the special case of Cauchy–Schwarz in the choice of the parameter $\gamma$ and the Rényi divergence is closely related to the Alpha divergence as shown in [26].

The Olivetti data set consists of intensity-value pictures of 40 individuals with small variations in viewpoint, large variation in expression and occasional addition of glasses. The data set contains 400 images (10 per person) of size $64 \times 64$. The COIL-20 data set contains images of 20 different objects viewed from 72 equally spaced orientations. In total we have 1440 images of $32 \times 32 = 1024$

[1] The Olivetti faces data set is publicly available from http://cs.nyu.edu/~roweis/data.html.

pixels. Like suggested in [16] we preprocessed the data by extracting the mean and reducing the dimension to 30 using PCA and successive transformation to unit variance features.

For the experiments we constructed 10 independent random initializations, which we reused in the algorithm with different divergences and values of the divergence parameter. To compare the different embeddings we use the one nearest neighbor classification error using the persons as labels. A quantitative evaluation based on the quality measure as proposed by [59,60] is included. Basically, this measure relies on $k$-intrusions and $k$-extrusions, which means it compares $k$-ary neighborhoods given in the original high-dimensional space with those occurring in the low dimensional space. Intrusions refer to samples intruding a neighborhood in the embedding space, while extrusions correspond to the number of samples

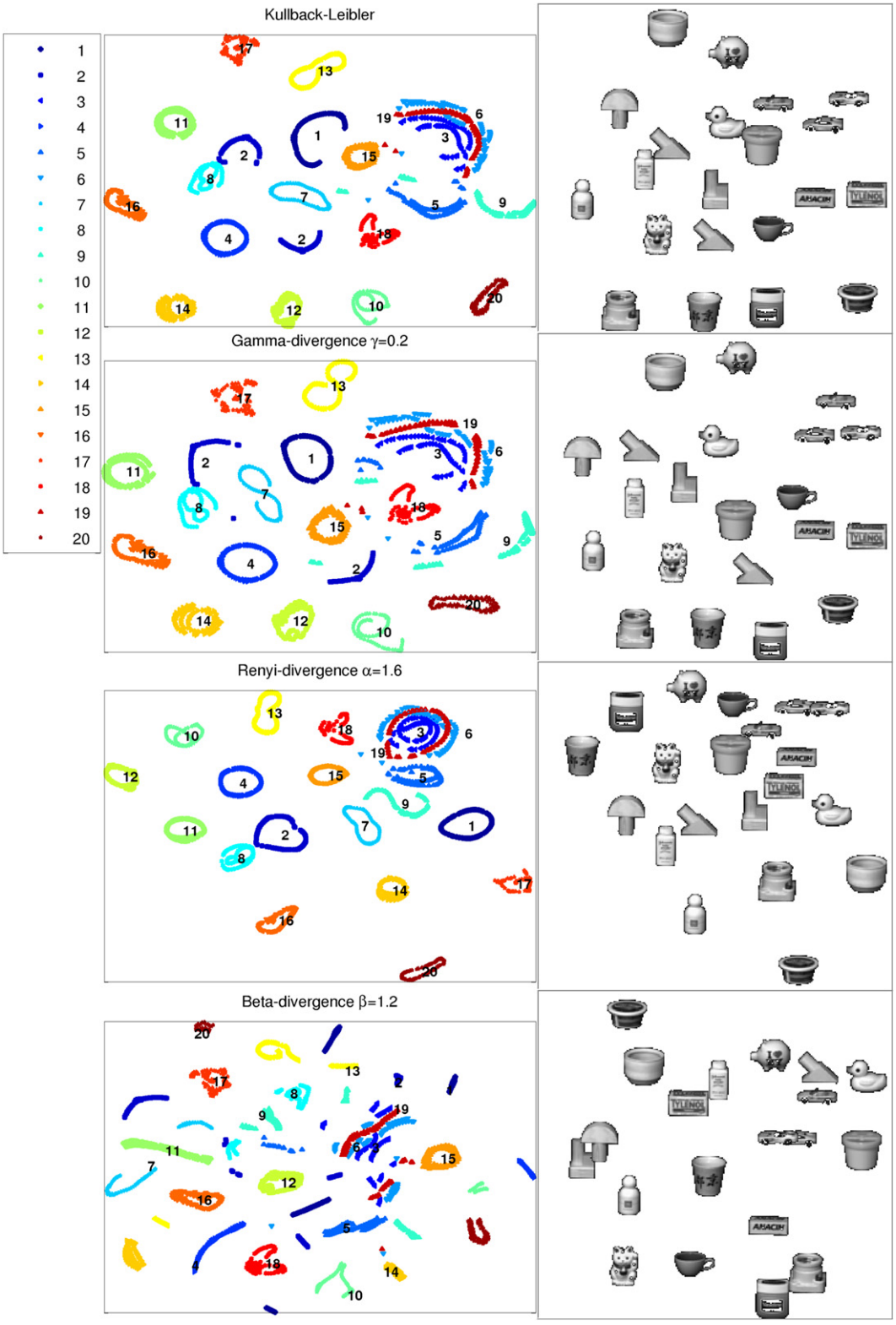


**Fig. 17.** Embeddings of the COIL-20 data set based on the same initialization for different divergences and perplexity 5.

which are missing in the projected $k$-ary neighborhoods. The overall quality $Q$ measures the percentage of data which is neither $k$-intrusive nor $k$-extrusive. In the optimal case all neighborhoods are exactly preserved, which results in a value of $Q=1$. The quantity $B$ measures the percentage of $k$-intrusions minus the percentage of $k$-extrusions in the projection and therefore shows the tendency of the mapping method: techniques with negative values for $B$ are characterized by extrusive behavior, while positive values display more intrusive behavior.

### 8.1. Olivetti faces

Fig. 12 shows the nearest neighbor errors of the embeddings of the Olivetti data as mean and standard deviation over the 10 random initializations for different perplexities. The parameter $\gamma$ of the Gamma divergences varies in the interval [0.2 2]. For Beta and Rényi the parameter ranges in the same interval excluding 1 and 2. Dependent on the perplexity the influence of the divergence varies. For small perplexities, greater values of $\gamma$ show better classification accuracy, while for large perplexities lower $\gamma$ yield better performance. The Gamma and Kullback–Leibler divergence show a quite robust behavior on this data set with respect to the parameter $\gamma$ and the perplexity. The quality and behavior of the Beta and Rényi divergence on the other hand vary a lot depending on the parameter and the actual perplexity. Also the variance with respect to the random initialization is much bigger for this data set using the Beta and Rényi divergence. The mean nearest neighbor error of the embedding is comparable to the other divergences if $\alpha=1.2$ and $\beta>1$ for all perplexities. For this data set the use of a different divergence leads to a slight improvement of the nearest neighbor classification compared to the Kullback–Leibler divergence and can be considered as an alternative measures.

Fig. 13 shows the quantitative evaluation on Olivetti using the intrusion- and extrusion measure mentioned above as mean over the 10 random initializations in the example case of perplexity 50. Again we observe small deviations in the behavior depending on the choice of the divergence. The Gamma divergence shows a little better quality for very small neighborhoods, while the Rényi divergence with $\alpha>1$ leads to a better quality for bigger neighborhoods. Some example visualizations are shown in Fig. 14. For comparison all visualizations are based on the same initialization.

### 8.2. COIL-20

Fig. 15 shows the nearest neighbor errors of the embeddings for COIL-20 as a mean and standard deviation over the 10 random initializations for different perplexities and Beta, Gamma and Rényi divergences with $\beta$, $\gamma$ and $\alpha$ varying in the interval [0.2 2]. Dependent on the perplexity the influence of the divergence varies. For small perplexities error free visualizations are possible in nearly all cases. Again the Beta and Rényi divergence shows quite different behavior dependent on the parameter. Nevertheless, for a small perplexity and $\{\beta,\alpha\}>1$ error free visualizations are also possible. For big perplexities in this data set the usage of the Gamma divergence leads to an improvement of the nearest neighbor classification in comparison with Kullback–Leibler. The Gamma divergence shows to be very robust to the actual choice of the perplexity.

Fig. 16 shows the quantitative evaluation using the intrusion- and extrusion measure mentioned above as mean over the 10 random initializations in the example case of perplexity 5. Again we observe deviations in the behavior dependent on the choice of the divergence. The Gamma divergence shows better quality for small neighborhoods, while using Rényi the quality for large neighborhoods can be improved if $\alpha>1$. The Beta divergence does not lead to

**Table 4**
Table of divergences and their t-SNE gradient.

| Divergence family | Functional gradient for t-SNE | Gradients for discrete data $\{x\}_{i=1}^n \in \mathbb{R}^N$ and $\{\xi\}_{i=1}^n \in \mathbb{R}^M$ |
|---|---|---|
| Kullback–Leibler equation (27) | $\frac{\partial D_{KL}}{\partial \xi} = 4\int \frac{\xi-\zeta}{1+r}(p-q)\,d\zeta$ | $\frac{\partial D_{KL}}{\partial \xi^i} = 4\sum_j^n \frac{\xi^i-\xi^j}{1+r_{\xi^i\xi^j}}(p_{x^ix^j}-q_{\xi^i\xi^j})$ |
| Itakura–Saito equation (28) | $\frac{\partial D_{IS}}{\partial \xi} = 4\int \frac{\xi-\zeta}{1+r}\left[\frac{p}{q}-1+q\int\left[1-\frac{p'}{q'}\right]\Pi_{r'}\,dr'\right]d\zeta$ | $\frac{\partial D_{IS}}{\partial \xi^i} = 4\sum_j^n \frac{\xi^i-\xi^j}{1+r_{\xi^i\xi^j}}\left[\frac{p_{x^ix^j}}{q_{\xi^i\xi^j}}-1+q_{\xi^i\xi^j}\sum_{kl}\left[1-\frac{p_{x^kx^l}}{q_{\xi^k\xi^l}}\right]\right]$ |
| Eta-divergence equation (30) | $\frac{\partial D_\eta}{\partial \xi} = 4(\eta^2-\eta)\int \frac{\xi-\zeta}{1+r}[(p-q)q^{(\eta-1)}-q \times \int[p'-q']q'^{(\eta-1)}\Pi_{r'}\,dr']\,d\zeta$ | $\frac{\partial D_\eta}{\partial \xi^i} = 4(\eta^2-\eta)\sum_j^n \frac{\xi^i-\xi^j}{1+r_{\xi^i\xi^j}}\left[(p_{x^ix^j}-q_{\xi^i\xi^j})q_{\xi^i\xi^j}^{(\eta-1)} -q_{\xi^i\xi^j}\cdot\sum_{kl}[p_{x^kx^l}-q_{\xi^k\xi^l}]q_{\xi^k\xi^l}^{(\eta-1)}\right]$ |
| Beta-divergence equation (32) | $\frac{\partial D_\beta}{\partial \xi} = 4\int \frac{\xi-\zeta}{1+r}[q^{\beta-1}(p-q)-q \times \int q'^{(\beta-1)}(p'-q')\Pi_{r'}\,dr']\,d\zeta$ | $\frac{\partial D_\beta}{\partial \xi^i} = 4\sum_j^n \frac{\xi^i-\xi^j}{1+r_{\xi^i\xi^j}}\left[q_{\xi^i\xi^j}^{\beta-1}(p_{x^ix^j}-q_{\xi^i\xi^j})-q_{\xi^i\xi^j} \times\sum_{kl}q_{\xi^k\xi^l}^{(\beta-1)}(p_{x^kx^l}-q_{\xi^k\xi^l})\right]$ |
| Alpha divergence equation (40) | $\frac{\partial D_\alpha}{\partial \xi} = \frac{4}{\alpha}\int \frac{\xi-\zeta}{1+r}(p^\alpha q^{1-\alpha}-q\int p'^\alpha q'^{(1-\alpha)}\Pi_{r'}\,dr')d\zeta$ | $\frac{\partial D_\alpha}{\partial \xi^i} = \frac{4}{\alpha}\sum_j^n \frac{\xi^i-\xi^j}{1+r_{\xi^i\xi^j}}\left[p_{x^ix^j}^\alpha q_{\xi^i\xi^j}^{1-\alpha}-q_{\xi^i\xi^j}\sum_{kl}p_{x^kx^l}^\alpha q_{\xi^k\xi^l}^{(1-\alpha)}\right]$ |
| Rényi divergence equation (43) | $\frac{\partial D_\alpha^R}{\partial \xi} = 4\int \frac{\xi-\zeta}{1+r}\left(\frac{p^\alpha q^{1-\alpha}}{\int p'^\alpha q'^{(1-\alpha)}\,dr'}-q\right)d\zeta$ | $\frac{\partial D_\alpha^R}{\partial \xi^i} = 4\sum_j^n \frac{\xi^i-\xi^j}{1+r_{\xi^i\xi^j}}\left[\frac{p_{x^ix^j}^\alpha q_{\xi^i\xi^j}^{1-\alpha}}{\sum_{kl}p_{x^kx^l}^\alpha q_{\xi^k\xi^l}^{(1-\alpha)}}-q_{\xi^i\xi^j}\right]$ |
| Tsallis divergence equation (44) | $\frac{\partial D_\alpha^T}{\partial \xi} = 4\int \frac{\xi-\zeta}{1+r}(p^\alpha q^{(1-\alpha)}-q\int p'^\alpha q'^{(1-\alpha)}\Pi_{r'}\,dr')d\zeta$ | $\frac{\partial D_\alpha^T}{\partial \xi^i} = 4\sum_j^n \frac{\xi^i-\xi^j}{1+r_{\xi^i\xi^j}}\left[p_{x^ix^j}^\alpha q_{\xi^i\xi^j}^{(1-\alpha)}-q_{\xi^i\xi^j}\sum_{kl}p_{x^kx^l}^\alpha q_{\xi^k\xi^l}^{(1-\alpha)}\right]$ |
| Hellinger divergence equation (45) | $\frac{\partial D_H}{\partial \xi} = 4\int \frac{\xi-\zeta}{1+r}(\sqrt{pq}-q\int\sqrt{p'q'}\Pi_{r'}\,dr')d\zeta$ | $\frac{\partial D_H}{\partial \xi^i} = 4\sum_j^n \frac{\xi^i-\xi^j}{1+r_{\xi^i\xi^j}}\left[\sqrt{p_{x^ix^j}q_{\xi^i\xi^j}}-q_{\xi^i\xi^j}\sum_{kl}\sqrt{p_{x^kx^l}q_{\xi^k\xi^l}}\right]$ |
| Gamma divergence equation (46) | $\frac{\partial D_\gamma}{\partial \xi} = 4\int \frac{\xi-\zeta}{1+r}\left(\frac{pq^\gamma}{\int p'q'^\gamma\,dr'}-\frac{q^{(\gamma+1)}}{\int q'^{(\gamma+1)}\,dr'}\right)d\zeta$ | $\frac{\partial D_\gamma}{\partial \xi^i} = 4\sum_j^n \frac{\xi^i-\xi^j}{1+r_{\xi^i\xi^j}}\left[\frac{p_{x^ix^j}q_{\xi^i\xi^j}^\gamma}{\sum_{kl}p_{x^kx^l}q_{\xi^k\xi^l}^\gamma}-\frac{q_{\xi^i\xi^j}^{(\gamma+1)}}{\sum_{kl}q_{\xi^k\xi^l}^{(\gamma+1)}}\right]$ |
| Cauchy–Schwarz equation (47) | $\frac{\partial D_{CS}}{\partial \xi} = 4\int \frac{\xi-\zeta}{1+r}\left(\frac{pq}{\int p'q'\,dr'}-\frac{q^2}{\int q'^2\,dr'}\right)d\zeta$ | $\frac{\partial D_{CS}}{\partial \xi^i} = 4\sum_j^n \frac{\xi^i-\xi^j}{1+r_{\xi^i\xi^j}}\left[\frac{p_{x^ix^j}q_{\xi^i\xi^j}}{\sum_{kl}p_{x^kx^l}q_{\xi^k\xi^l}}-\frac{q_{\xi^i\xi^j}^2}{\sum_{kl}q_{\xi^k\xi^l}^2}\right]$ |

an improvement for this particular data set and although error free visualizations are possible it is not satisfying, because it resembles only very close neighborhoods but scatters the trajectories. Some error free example visualizations are shown in Fig. 17. For comparison all visualizations are based on the same initialization. Note that, for example, the data points representing object 1 are chained on a bended line using the Kullback–Leibler divergence, while it is visualized in a closed loop using the Gamma divergence. Interestingly the use of the Rényi divergence with $\alpha = 1.6$ resembles the desired loop structure for nearly all objects. The cars (objects 3, 6 and 19) are visualized as rings not as long bands as seen using Kullback–Leibler and the Gamma divergence. Also objects 2 and 9 are no longer divided into pieces but a connected structure. Besides some topologic defects and the non-closed loop of object 9 the visualization using Rényi is a quite good estimate of the optimal visualization one would expect for this particular data set.

## 9. Conclusion and outlook

The original SNE and t-SNE formulation employ the Kullback–Leibler divergence. In this contribution we provide a mathematical foundation for the use of arbitrary divergences and their derivatives such that they can immediately be plugged into the existing algorithms. We provide the reader with alternative measures, which can be used if the results using Kullback–Leibler are not satisfying.

For this purpose we characterized main subclasses of divergences following [26]: Bregman-, *f*- and Gamma divergences. We used the mathematical framework of Fréchet derivatives to derive the gradients for a wide range of important divergences as summarized in Table 4.

We studied the behavior of the divergences in some experiments inspired by image processing. From the experiments it is clearly visible that the divergences show different behavior for different problems. Although we are not yet able to deliver an overall recipe for choosing a particular divergence in a given task, we can still argue that it might be advantageous to try alternative measures if the results are not satisfying. We demonstrate the use of divergences taken from all three main families on two example data sets, namely the Olivetti faces and COIL-20 data set. Performances are compared in terms of the nearest neighbor classification error of the embeddings, the quality as measured by intrusion- and extrusion behavior [59,60] and by visual inspection. The Gamma divergence shows quite robust behavior with respect to the parameter $\gamma$ and the perplexity used. For the Beta and Rényi divergence on the other hand the behavior varies a lot with the actual choice of $\beta$, $\alpha$ and the perplexity. Nevertheless, using Rényi we can observe an improvement of the global quality looking at bigger neighborhoods in the Experiments. Especially the visualization of the COIL-20 data resembles very nicely the structure of the data set. Data sets in which a good mapping of bigger neighborhoods is desired the use of the Rényi divergence based potential for improvement.

The investigation of further divergences on more data sets will be addressed in further studies. Furthermore divergences like Alpha-, Beta-, Eta-, Gamma-, generalized Rényi, and generalized Kullback–Leibler divergence do not require probability densities as inputs, but can be applied to positive measures. Through normalization information might get lost, so the use of generalized divergences on non-normalized neighborhood functions improves performances, potentially, and will be investigated in forthcoming projects.

## Acknowledgments

This work was supported by the "Nederlandse Organisatie voor Wetenschappelijk Onderzoek (NWO)" under project code 612.066.620 and by the "German Science Foundation (DFG)". Further, financial support from the Cluster of Excellence 277 Cognitive Interaction Technology funded in the framework of the German Excellence Initiative is gratefully acknowledged.

## References

[1] P.L. Lai, C. Fyfe, Bregman divergences and multi-dimensional scaling, in: 15th International Conference on Neuro-Information Processing: (ICONIP), Revised Selected Papers, Part II, Springer-Verlag, Auckland, New Zealand, 2008, pp. 935–942. doi:10.1007/978-3-642-03040-6_114.
[2] J.B. Tenenbaum, V.d. Silva, J.C. Langford, A global geometric framework for nonlinear dimensionality reduction, Science 290 (5500) (2000) 2319–2323. http://dx.doi.org/:10.1126/science.290.5500.2319.
[3] V. De Silva, J.B. Tenenbaum, Global versus local methods in nonlinear dimensionality reduction, in: Advances in Neural Information Processing Systems, vol. 15, 2003, pp. 705–712. doi:10.1.1.9.3407.
[4] S.T. Roweis, L.K. Saul, Nonlinear dimensionality reduction by locally linear embedding, Science 290 (5500) (2000) 2323–2326. http://dx.doi.org/:10.1126/science.290.5500.2323.
[5] L.J.P. van der Maaten, E.O. Postma, H.J. van den Herik, Dimensionality Reduction: A Comparative Review, Technical Report TiCC-TR 2009-005, Tilburg University, October 2009.
[6] M. Brand, Charting a Manifold, Technical Report 15, Mitsubishi Electric Research Laboratories (MERL), 2003.
[7] Y. Teh, S. Roweis, Automatic alignment of local representations, in: Advances in Neural Information Processing Systems, vol. 15, 2003, pp. 841–848.
[8] K. Fukunaga, Introduction to Statistical Pattern Recognition, Computer Science and Scientific Computing Series, 2nd ed., Academic Press, 1990.
[9] G. Baudat, F. Anouar, Generalized discriminant analysis using a kernel approach, Neural Comput. 12 (10) (2000) 2385–2404.
[10] T. Kohonen, Self-Organizing Maps, 3rd ed., Springer, Berlin, Heidelberg, New York, 2001.
[11] T. Villmann, B. Hammer, F.-M. Schleif, T. Geweniger, W. Hermann, Fuzzy classification by fuzzy labeled neural gas, Neural Networks 19 (6–7) (2006) 772–779.
[12] K. Bunte, P. Schneider, B. Hammer, F.-M. Schleif, T. Villmann, M. Biehl, Discriminative Visualization by Limited Rank Matrix Learning, Technical Report MLR-03-2008, Leipzig University, 2008.
[13] K. Bunte, B. Hammer, A. Wismüller, M. Biehl, Adaptive local dissimilarity measures for discriminative dimension reduction of labeled data, Neurocomputing 73 (7–9) (2010) 1074–1092. http://dx.doi.org/:10.1016/j.neucom.2009.11.017.
[14] J. Lee, M. Verleysen, Nonlinear Dimensionality Reduction, 1st ed., Springer, 2007.
[15] G. Hinton, S. Roweis, Stochastic neighbor embedding, in: Advances in Neural Information Processing Systems, vol. 15, MIT Press, 2003, pp. 833–840.
[16] L. van der Maaten, G. Hinton, Visualizing data using t-SNE, J. Mach. Learn. Res. 9 (2008) 2579–2605.
[17] J.A. Lee, M. Verleysen, Shift-invariant similarities circumvent distance concentration in stochastic neighbor embedding and variants, Procedia Comput. Sci. 4 (2011) 538–547.
[18] F. Rossi, N. Delannay, B. Conan-Guez, M. Verleysen, Representation of functional data in neural networks, Neurocomputing 64 (2005) 183–210.
[19] J.A. Lee, M. Verleysen, Generalization of the $l_p$ norm for time series and its application to self-organizing maps, in: M. Cottrell (Ed.), Proceedings of Workshop on Self-Organizing Maps (WSOM), Paris, Sorbonne, 2005, pp. 733–740.
[20] J. Ramsay, B. Silverman, Functional Data Analysis, 2nd ed., Springer, New York, 2006.
[21] T. Villmann, Sobolev Metrics for Learning of Functional Data—Mathematical and Theoretical Aspects, Machine Learning Reports 1 (MLR-03-2007), 2007, pp. 1–15.
[22] T. Villmann, F.-M. Schleif, Functional vector quantization by neural maps, in: J. Chanussot (Ed.), Proceedings of First Workshop on Hyperspectral Image and Signal Processing: Evolution in Remote Sensing (WHISPERS 2009), IEEE Press, 2009, pp. 1–4.
[23] T. Villmann, S. Haase, Divergence based vector quantization, Neural Comput. 23 (5) (2011) 1343–1392.
[24] E. Mwebaze, P. Schneider, F.M. Schleif, J.R. Aduwo, J.A. Quinn, S. Haase, T. Villmann, M. Biehl, Divergence-based classification in learning vector quantization, Neurocomputing 74 (2011) 1429–1435. http://dx.doi.org/:10.1016/j.neucom.2010.10.016.
[25] K. Bunte, F.-M. Schleif, S. Haase, T. Villmann, Mathematical foundations of the self organized neighbor embedding (SONE) for dimension reduction and visualization, in: M. Verleysen (Ed.), Proceedings of European Symposium on Artificial Neural Networks (ESANN), Bruges, Belgium, 2011, pp. 29–34.
[26] A. Cichocki, R. Zdunek, A. Phan, S.I. Amari, Non-negative Matrix and Tensor Factorizations, Wiley, Chichester, Hoboken, NJ, 2009.
[27] B.A. Frigyik, S. Srivastava, M. Gupta, An Introduction to Functional Derivatives, Technical Report UWEETR-2008-0001, Department of Electrical Engineering, University of Washington, Seattle, 2008.

[28] I. Kantorowitsch, G. Akilow, Funktionalanalysis in normierten Räumen, 2nd ed., Akademie-Verlag, Berlin, 1978.
[29] A. Cichocki, S.-I. Amari, Families of alpha- beta- and gamma-divergences: flexible and robust measures of similarities, Entropy 13 (2010) 134–170.
[30] A. Banerjee, S. Merugu, I.S. Dhillon, J. Ghosh, Clustering with Bregman divergences, J. Mach. Learn. Res. 6 (2005) 1705–1749.
[31] A. Banerjee, I. Dhillon, J. Ghosh, S. Merugu, D.S. Modha, A generalized maximum entropy approach to Bregman co-clustering and matrix approximation, in: KDD '04: Proceedings of the 10th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, ACM, New York, NY, USA, 2004, pp. 509–514. doi:10.1145/1014052.1014111.
[32] L.M. Bregman, The relaxation method of finding the common point of convex sets and its application to the solution of problems in convex programming, USSR Comput. Math. Math. Phys. 7 (1967) 200–217.
[33] I.S. Dhillon, S. Sra, Generalized nonnegative matrix approximations with Bregman divergences, in: Neural Information Processing Systems, Vancouver, Canada, 2005, pp. 283–290. doi:10.1.1.72.5975.
[34] I.S. Dhillon, J.A. Tropp, Matrix nearness problems with Bregman divergences, SIAM J. Matrix Anal. Appl. 29 (4) (2007) 1120–1146. http://dx.doi.org/:10.1137/060649021.
[35] N. Murata, T. Takenouchi, T. Kanamori, Information geometry of U-Boost and Bregman divergence, Neural Comput. 16 (2004) 1437–1481.
[36] S. Eguchi, Y. Kano, Robustifying Maximum Likelihood Estimation, Technical Report 802, Tokyo-Institute of Statistical Mathematics, Tokyo, 2001.
[37] S. Kullback, R.A. Leibler, On information and sufficiency, Ann. Math. Stat. 22 (1951) 49–86.
[38] J.N. Kapur, Measures of Information and Their Applications, Wiley-Interscience, Hoboken, NJ, 1994.
[39] C.E. Shannon, A Mathematical Theory of Communication, CSLI Publications, 1948.
[40] F. Itakura, S. Saito, Analysis synthesis telephony based upon the maximum likelihood method, in: Independent Component Analysis, 1968.
[41] N. Bertin, C. Fevotte, R. Badeau, A tempering approach for Itakura–Saito non-negative matrix factorization. With application to music transcription, in: ICASSP '09: Proceedings of the 2009 IEEE International Conference on Acoustics, Speech and Signal Processing, IEEE Computer Society, Washington, DC, USA, 2009, pp. 1545–1548. doi:10.1109/ICASSP.2009.4959891.
[42] F. Nielsen, R. Nock, Sided and symmetrized Bregman centroids, IEEE Trans. Inf. Theory 55 (2009) 2882–2904. http://dx.doi.org/:10.1109/TIT.2009.2018176.
[43] A. Basu, N.L.H. Ian, R. Harris, M.C. Jones, Robust and efficient estimation by minimising a density power divergence, Biometrika 85 (3) (1998) 549–559. http://dx.doi.org/:10.1093/biomet/85.3.549.
[44] M. Mihoko, S. Eguchi, Robust blind source separation by beta divergence, Neural Comput. 14 (8) (2002) 1859–1886. http://dx.doi.org/:10.1162/089976602760128045.
[45] I. Csiszár, Information-type measures of difference of probability distributions and indirect observations, in: Studia Scientiarum Mathematicarum Hungarica, vol. 2, 1967, pp. 299–318.
[46] I. Csiszár, A class of measures of informativity of observation channels, in: Periodica Mathematica Hungarica, vol. 2, 1972, pp. 191–213.
[47] S.I. Amari, H. Nagaoka, Methods of information geometry, in: Translations of Mathematical Monographs, vol. 191, Oxford University Press, New York, 2000.
[48] I.J. Taneja, P. Kumar, Relative information of type s, Csiszár's f-divergence, and information inequalities, Inf. Sci. 166 (1–4) (2004) 105–125.
[49] F. Österreicher, Csiszár f-divergences—Basic Properties, Technical Report, Research Report Collection, 2002.
[50] F. Liese, I. Vajda, Convex statistical distances, in: Teubner-Texte zur Mathematik, vol. 95, Teubner-Verlag, Leipzig, 1987.
[51] S.-I. Amari, Differential–geometrical Methods in Statistics, Springer, Berlin, 1985.
[52] A. Rényi, On measures of entropy and information, in: Proceedings of the 4th Berkeley Symposium on Mathematical Statistics and Probability, vol. 1, University of California Press, Berkeley, 1961, p. 547.
[53] A. Rényi, Probability Theory, North-Holland Series in Applied Mathematics and Mechanics, vol. 10, Amsterdam, 1970.
[54] H. Fujisawa, S. Eguchi, Robust parameter estimation with a small bias against heavy contamination, Multivariate Anal. 99 (9) (2008) 2053–2081. http://dx.doi.org/:10.1016/j.jmva.2008.02.004.
[55] J.C. Principe, D. Xu, J.W. Fisher III, Information-theoretic learning, in: S. Haykin (Ed.), Unsupervised Adaptive Filtering, 2nd ed., vol. 1, Wiley, New York, 2000 (Chapter 7).
[56] R. Jenssen, An information theoretic approach to machine learning, PhD dissertation, University of Tromsø, Department of Physics, 2005.
[57] R. Jenssen, J.C. Principe, D. Erdogmus, T. Eltoft, The Cauchy–Schwarz divergence and Parzen windowing: connections to graph theory and Mercer kernels, J. Franklin Inst. 343 (6) (2006) 614–629. http://dx.doi.org/:10.1016/j.jfranklin.2006.03.018.
[58] S.A. Nene, S.K. Nayar, H. Murase, Columbia Object Image Library (COIL-20), Technical Report CUCS-005-96, Columbia University, 1996.
[59] J.A. Lee, M. Verleysen, Rank-based quality assessment of nonlinear dimensionality reduction, in: 16th European Symposium on Artificial Neural Networks (ESANN), Bruges, Belgium, 2008, pp. 49–54.
[60] J.A. Lee, M. Verleysen, Quality assessment of dimensionality reduction: rank-based criteria, Neurocomputing 72 (7–9) (2009) 1431–1443. http://dx.doi.org/:10.1016/j.neucom.2008.12.017.

**Kerstin Bunte** graduated at the Faculty of Technology at the University of Bielefeld, Germany, and joined the Institute of Mathematics and Computing Science of the University of Groningen, The Netherlands, in September 2007. She received a Ph.D. in Computer Science in December 2011. Her recent work has focused on Machine Learning techniques, especially Learning Vector Quantization and their usability in the field of image processing, supervised dimension reduction and visualization. Further information can be obtained from http://www.cs.rug.nl/~kbunte/.

**Sven Haase** received his Diploma in Applied Mathematics from the University of Applied Sciences Mittweida, Germany, in 2009. Currently, he is a Ph.D. candidate in the Computational Intelligence Group of the Faculty Mathematics, Sciences and Computer Science at the University of Applied Sciences Mittweida. His research interest is in Machine Learning, especially theoretical aspects of prototype-based classification methods.

**Michael Biehl** received a Ph.D. in Theoretical Physics from the University of Giessen, Germany, in 1992 and the venia legendi in Theoretical Physics from the University of Würzburg, Germany, in 1996. He is currently Associate Professor with Tenure in Computing Science at the University of Groningen, The Netherlands. His main research interest is in the theory, modelling and application of Machine Learning techniques. He is furthermore active in the modelling and simulation of complex physical systems. He has co-authored more than 100 publications in international journals and conferences; preprint versions and further information can be obtained from http://www.cs.rug.nl/~biehl/.

**Thomas Villmann** holds a diploma degree in Mathematics and received his Ph.D. in Computer Science in 1996 and his venia legendi in the same subject in 2005, both from the University of Leipzig. From 1997 to 2009 he led the research group of computational intelligence of the clinic for psychotherapy at Leipzig University. Since 2009 he is a full professor for Technomathematics and Computational Intelligence at the University of Applied Science Mittweida, Germany. He is founding member of the German chapter of ENNS, member of the IEEE TaskforceCIS task force on Data Visualization and Data Analysis, and speaker of the working group "neural Networks" of the German Computer Science Society (GI). His research areas include a broad range of machine learning approaches like neural maps, clustering, classification, pattern recognition and evolutionary algorithms as well as applications in medicine, bioinformatics, satellite remote sensing and others.